International Journal of Interactive Multimedia and Artificial Intelligence

June 2015, Vol III, Number 3, ISSN: 1989-1660

The question of whether a computer can think is no more interesting than the question of whether a submarine can swim.

Edsger W. Dijkstra

http://www.ijimai.org

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE AND INTERACTIVE MULTIMEDIA
ISSN: 1989-1660 - VOL. III, NUMBER 3

Editor's Note

The International Journal of Interactive Multimedia and Artificial Intelligence provides an interdisciplinary forum in which scientists and professionals can share their research results and report new advances on Artificial Intelligence and Interactive Multimedia techniques.

The research works presented in this issue are based on various topics of interest, among which are included: 3D Interface, Information Extraction, Artificial immune system, Security in Cloud Computing, Genetic Algorithm, Activity recognition, 3D Render Distribution, Software visualization, Event Perception, New Programming Language, Distributed computing, MOOC environments, etc.

Contreras, D. Et al. [1] talks about a novel framework to tightly integrate interactive recommendation systems in a 3D virtual environment. Specifically, we propose to integrate a Collaborative Conversational Recommender (CCR) in a 3D social virtual world. Our CCR Framework defines three layers: the user interaction layer (3D Collaborative Space Client), the communication layer (3D Collaborative Space Server), and the recommendation layer (Collaborative Conversational Recommender). Additionally, we evaluate the framework based on several usability criteria such as learnability, perceived efficiency and effectiveness. Results demonstrate that users positively valued the experience.

Ghoulam A. Et al. [2] presents a local grammar approach to extract medical named entities from French patient clinical reports. Experimental results show that the proposed approach achieved an F-Measure of 90. 06%.

Baldominos, A. G. Et al. [3] writes about a design of an evolutionary algorithm for building classifiers specifically aimed towards performing classification and sentiment analysis over texts. Moreover, it has properties taken from Artificial Immune Systems, as it tries to resemble biological systems since they are able to discriminate harmful from innocuous bodies (in this case, the analogy could be established with negative and positive texts respectively). A framework, namely OpinAIS, is developed around the evolutionary algorithm, which makes it possible to distribute it as an open-source tool, which enables the scientific community both to extend it and improve it. The framework is evaluated with two different public datasets, the first involving voting records for the US Congress and the second consisting in a Twitter corpus with tweets about different technology brands, which can be polarized either towards positive or negative feelings; comparing the results with alternative machine learning techniques and concluding with encouraging results. Additionally, as the framework is publicly available for download, researchers can replicate the experiments from this paper or propose new ones.

Venkateshwaran K., D. Et al. [4] talk about that Agent can play a key role in bringing suitable cloud services to the customer based on their requirements. In agent based cloud computing, agent does negotiation, coordination, cooperation and collaboration on behalf of the customer to make the decisions in efficient manner. However the agent based cloud computing have some security issues like (a.) addition of malicious agent in the cloud environment which could demolish the process by attacking other agents, (b.) denial of service by creating flooding attacks on other involved agents. (c.) Some of the exceptions in the agent interaction protocol such as Not-Understood and Cancel_Meta protocol can be misused and may lead to terminating the connection of all the other agents participating in the negotiating services. Also, this paper proposes algorithms to solve these issues to ensure that there will be no intervention of any malicious activities during the agent interaction.

Zaldaña, H. Et al. [5] shows an alternative to this process is the UV disinfection of drinking water. Most of the devices in the market use UV bulbs or mercury lamps. The UV LED, which is cheaper and smaller, allows creating new smaller devices. The main contribution of this paper is the use of Genetic Algorithms to help design a drinking water device with UV LEDs.

Pekka, S. Et al. [6] present a short abstract of your PhD thesis, called "Recognizing Human Activities Based on Wearable Inertial Measurements - Methods and Applications.

Bolívar, H. D. Et al. [7] talk about a research that is based on the method 'divide and rule', that is, volumetric surfaces are subdivided using a tree-KD of sequence of scenes in a game, so reducing the surface into small sets of points. Reconstruction efficiency is improved, because the search of data is performed in local and small regions. Processes are modeled through a finite set of states that are built using Hidden Markov Models with domains configured by heuristics. Six test that control the states of each heuristic, including the number of intervals are carried out to validate the proposed model. This validation concludes that the proposed model optimizes response frames per second, in a sequence of interactions.

Almeida-Martínez, F. J. Et al. [8] shows a reviews software visualization focused on the educational environment. Software visualization is a very wide study field, so we have focused on two areas: recursion visualization and parsers' visualization. The paper contains a retrospective about what has been made on it, what lacks we have found and the solution provided by the authors: SRec and VAST, two software tools trying to make a significant difference between them and the software made before.

Jain, S. Et al. [9] propose an extension of perception process in an existing emotion model, EMIA and suggest the formalization of event perception and appraisal processes to make it adaptable. This has been carried out using five parameters for event description along-with fuzzy logic which makes the process more effective yet simple.

González, C. Et al. [10] shows a comparative the new programming language of Apple, Swift, with the main

DOI: 10.9781/ijimai.2015.330

programming language of Apple before Swift, Objective-C. They are going to show the differences, characteristics and novelties to verify the words of Apple about Swift. With that they want to answer the next question: Is Swift a new programming language easier, more secure and quicker to develop than Objective-C?

Torres, E. Et al. [11] expose a report with a solution based on dynamic partial replication where the number of replicas for each file and its management is handled by an agent architecture. they compare their solution with full replication and with static partial replication both in terms of storage capacity consumption and service time. Their results show that their proposed solution provides equivalent performance with a better use of disk storage capacity.

Corbi, A. D. Et al. [12] shows a software framework that aims at assisting teachers in MOOCs during math-nature exercise correction tasks. This framework might fit for math, physics, or technical teachings. As a test experience, they apply it to 300+ physics homework bulletins from 80+ students. Test results show their solution might turn very useful when guiding assistant teachers during correction shifts.

Rios, S. Et al [13] explains how wereable technologies are capable of measuring the heart beat and, further, using other sensors like Accelerometer and Gyroscope, embedded on a simple clock allow us to monitor the physical activity of the user. Their main goal is to use the pulsations measurements in conjunction with the physical activity for the detection of driver drowsiness/sleepiness in advance in order to prevent accidents derived from fatigue.

Dr. Carlos Montenegro

REFERENCES

[1] Contreras, D. Et al. "A 3D Visual Interface for Critiquing-based Recommenders: Architecture and Interaction", International Journal of Artificial Intelligence and Interactive Multimedia, vol. 3, no. 3, pp. 7-15, 2015.

[2] Ghoulam A. Et al. "Using Local Grammar for Entity Extraction from Clinical Reports", International Journal of Artificial Intelligence and Interactive Multimedia, vol. 3, no. 3, pp. 16-24, 2015.

[3] Baldominos, A. G. Et al. "OpinAIS: An Artificial Immune System-based Framework for Opinion Mining", International Journal of Artificial Intelligence and Interactive Multimedia, vol. 3, no. 3, pp. 25-34, 2015.

[4] Venkateshwaran K., D. Et al. "Security Framework for Agent-Based Cloud Computing", International Journal of Artificial Intelligence and Interactive Multimedia, vol. 3, no. 3, pp. 35-42, 2015.

[5] Zaldaña, H. Et al. "The Use of Genetic Algorithms in UV Disinfection of Drinking Water", International Journal of Artificial Intelligence and Interactive Multimedia, vol. 3, no. 3, pp. 43-48, 2015.

[6] Pekka, S. Et al. "Recognizing Human Activities Based on Wearable Inertial Measurements - Methods and Applications", International Journal of Artificial Intelligence and Interactive Multimedia, vol. 3, no. 3, pp. 49-50, 2015.

[7] Bolívar, H. D. Et al. "An Architecture Approach for 3D Render Distribution using Mobile Devices in Real Time", International Journal of Artificial Intelligence and Interactive Multimedia, vol. 3, no. 3, pp. 51-60, 2015.

[8] Almeida-Martínez, F. J. Et al. "SRec and VAST: Visualizing Software with a Student-Centered Aim", International Journal of Artificial Intelligence and Interactive Multimedia, vol. 3, no. 3, pp. 61-68, 2015.

[9] Jain, S. Et al. "Formalization of Event Perception and Event Appraisal Process", International Journal of Artificial Intelligence and Interactive Multimedia, vol. 3, no. 3, pp. 69-73, 2015.

[10] González, C. Et al. "Swift vs. Objective-C: A New Programming Language", International Journal of Artificial Intelligence and Interactive Multimedia, vol. 3, no. 3, pp. 74-81, 2015.

[11] Torres, E. Et al. "A Quantitative Justification to Dynamic Partial Replication of Web Contents through an Agent Architecture", International Journal of Artificial Intelligence and Interactive Multimedia, vol. 3, no. 3, pp. 82-88, 2015.

[12] Corbi, A. D. Et al. "Semi-automated correction tools for math-like exercises in MOOC environments", International Journal of Artificial Intelligence and Interactive Multimedia, vol. 3, no. 3, pp. 89-95, 2015.

[13] Rios, S. Et al." Variation of the heartbeat and activity as an indicator of drowsiness at the wheel using a smartwatch", International Journal of Artificial Intelligence and Interactive Multimedia, vol. 3, no. 3, pp. 96-100, 2015

TABLE OF CONTENTS

OPEN ACCESS JOURNAL

ISSN: 1989-1660

COPYRIGHT NOTICE

A 3D Visual Interface for Critiquing-based Recommenders: Architecture and Interaction

David Contreras[1,2], Maria Salamó[2], Inmaculada Rodríguez[2], Anna Puig[2]

[1]*Facultad de Ingeniería y Arquitectura, Universidad Arturo Prat, Chile*
[2]*Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Spain*

***Abstract* — Nowadays e-commerce websites offer users such a huge amount of products, which far from facilitating the buying process, actually make it more difficult. Hence, recommenders, which learn from users' preferences, are consolidating as valuable instruments to enhance the buying process in the 2D Web. Indeed, 3D virtual environments are an alternative interface for recommenders. They provide the user with an immersive 3D social experience, enabling a richer visualisation and increasing the interaction possibilities with other users and with the recommender. In this paper, we focus on a novel framework to tightly integrate interactive recommendation systems in a 3D virtual environment. Specifically, we propose to integrate a Collaborative Conversational Recommender (CCR) in a 3D social virtual world. Our CCR Framework defines three layers: the user interaction layer (3D Collaborative Space Client), the communication layer (3D Collaborative Space Server), and the recommendation layer (Collaborative Conversational Recommender). Additionally, we evaluate the framework based on several usability criteria such as learnability, perceived efficiency and effectiveness. Results demonstrate that users positively valued the experience.**

***Keywords* — Collaborative Conversational Recommenders, Intelligent Collaborative 3D Interface, 3D Virtual Worlds.**

I. INTRODUCTION

E-COMMERCE customers are used to being guided by some type of e-assistant which helps them with information overload. Recommendation web-engines assist the user in a variety of e-commerce applications, such as those for buying music, books and mobile phones. Recommender systems typically endorse a list of suggestions close to the user preferences, through collaborative or content-based filtering. Collaborative filtering methods use a model from the past behavior of the user as well as selections made by other people before. Content-based filtering use a description of the product and a profile of the user's interests. Conversational recommenders is a form of content-based filtering that is well suited to many product recommendations.

Critiquing-based recommender systems are a kind of interactive Conversational[1] recommenders which help users to navigate through a product space, alternatively making product suggestions and eliciting user feedback in the form of critiques such as "I would like something cheaper" or "with faster processor speed" [1],[5]. Thus, a critique is a directional feature preference in relation to the current recommendation. Collaborative Conversational Recommenders (CCR) [6] exploit not only the critique-based feedback of the user, but also the on-line collaboration between users who have similar or different goals.

A typical session with a conversational recommender consists of a series of recommend-review-revise-update cycles [14]. First, in each cycle a new product is *recommended* to the user based on the current product query. Second, the user provides some form of feedback (*review*) regarding the suitability of the item. Third, the query is *revised* for the next cycle. Finally, the user model is *updated* by adding the last critique and pruning all the critiques that are inconsistent with it. The user finishes the process when she finds a suitable product or give the recommendation process up.

Recommendation engines are widespread in the (2D) web where they provide an interactive experience to users, and where users may collaborate using chat, messages and social networks. However, in general, the on-line collaboration among users in e-commerce (2D) web-based spaces is constrained to chat interaction, loosing some sensory information of other people, such as the physical behavior, gestures as well as voice-based and text-based chat. Nevertheless, 3D interfaces further facilitate the collaboration among users, who feel immersed and have better social interaction with each other [11]. Additionally, virtual agents can provide personalized services, and then increasing consumer trust in e-commerce [8].

Furthermore, although an initial goal of 3D business virtual worlds has been to generate brand awareness and increase traffic to 2D e-commerce web pages, virtual store customers generally want to stay inside the 3D virtual world and thereby complete the buying process within the same virtual

[1] In this work the term "Conversational" is employed in the scope of the Recommender Systems, but it does not refer to a typical user action inside of a 3D virtual environment such as text or voice chats.

DOI: 10.9781/ijimai.2015.331

environment [12], where can share their experiences with others. However, little research has been done on the integration of recommender systems in 3D virtual environments. Recent works integrate collaborative filtering approaches in virtual worlds [17][18] but, as far as we know, no one exploits the benefits of a conversational recommender.

In this paper, we present a novel framework to integrate recommendation systems in 3D virtual environments. Specifically, we propose to integrate a Collaborative Conversational Recommender (CCR) in a 3D social virtual world to provide the users with an immersive and collaborative recommendation experience. Note that this is not a group recommender but a conversational recommender which allows the user to interact with it either in a individually or collaborative way. This framework is composed by three layers: the interface layer (3D virtual world), the communication layer and the recommender layer. Interactions between the user layer and the recommender layer flow as messages across the communication layer. Moreover, this framework deals with both individual and collaborative critiques. Finally, we evaluate the approach by means of different usability criteria such as learnability, satisfaction, effectiveness, and both perceived and real efficiency.

The remaining of this paper is structured as follows: Section 2 presents related work; Section 3 describes in depth the proposal; Section 4 evaluates the performance and the usability of our proposal with real-users; Finally, Section 5 concludes the paper.

II. RELATED WORK

The majority of conversational recommenders that use critiquing as feedback mechanism assume web-based platforms, for example QuickShop [14] or Example Critiquing [13]. A different approach is used by CATS [10], which defines a group recommender that uses an interactive table-top device for allowing the interaction of multiple users through a touch screen. There are also proposals on mobile devices, such as MobyRek [15] and CritiqueShop [19].

Relatively little research have been focused on exploiting 3D interfaces for recommender systems. In this way, Second Life[2] has been used by some researchers as an interactive recommendation platform [2],[17]. EEG system [7] used OpenSimulator, an open-source 3D virtual world platform that follows Second Life protocols so that users connect using any Second Life viewer. Other works have used more specific tools to develop recommenders in 3D virtual environments, such as Java3D with VRML (Virtual Reality Markup Language) [18] and [9].

In relation to the application domain of previous works, most of them have been focused on implementing shopping assistants. For example, Xu and Yu [18] present a solution of virtual shopping mall on the Internet through a recommender based on data mining technologies. Other authors recommend virtual objects inside a virtual reality interface, using a classification method based on off-line historical data [9]. Others have focused on recommending locations (places inside the virtual world) made by other users [17]. Recently, [7] have proposed to evaluate pre-purchase ratings (in addition to traditional post-purchase ratings in recommendation process), which were based on electroencephalogram (EEG) signals obtaining the users' positive emotions while interacting with virtual products before to purchase.

Regarding the recommendation method, the majority of previous studies have used a traditional Collaborative Filtering (CF) method [2], [7], [17], [18], [20] for generating users recommendations. CF [4] is based on historical data and does not necessarily imply a direct on-line interaction among users. There is also a hybrid approach that is based on both collaborative-filtering and content-based methods [9]. However, this hybrid approach does not allow on-line user collaboration.

Our proposal uses a 3D collaborative platform that provide users with meaningful visualizations, interaction mechanisms and a great sense of immersion. Specifically, we use OpenSimulator server platform, with a Second Life Viewer client, allowing a better separation between the different functionalities of the 3D virtual environment and the recommender system. Our CCR framework focuses on e-commerce applications, which perform massive products recommendation (e.g. smartphone, pc, travel). Moreover, we have enriched the traditional critiquing feedback mechanism allowing a collaborative on-line selection of products among users. The main advantage of the CCR framework with respect to previous studies is that none of them is conversational nor exploits the collaborative features of 3D virtual environments. In addition, we propose a novel framework that can be used to develop new solutions for integrating recommenders in 3D virtual environments.

III. PROPOSAL

This section presents our Collaborative Conversational Recommender (CCR) framework, which integrates a Conversational Recommender in a 3D interface to provide users with an immersive and collaborative recommendation experience.

Figure 1 shows the three layers in the CCR framework. In the top of the figure, the *3D Collaborative Space Client* is an immersive 3D virtual space where users interact each other and with the recommender to acquire a desired product. In the bottom, the *Collaborative Conversational Recommender* layer hosts recommendation algorithms, case bases (*CB*) and users models needed for the recommendation process. In the middle, the *3D Collaborative Space Server* is the communication layer responsible for the connection between previous –interface and recommender– layers, as well as for users and 3D content management.

As previously introduced, a typical session with a conversational (interactive) recommender consists of a series of recommend-review-revise-update cycles, where both the user and the recommender interact each other in several cycles. A recommendation cycle starts when the recommender (in the CCR layer) provides the user (in the *3D Collaborative*

[2] It is a massively online 3D content-based virtual world that permits users to construct, interact, and inhabit their own 3D world.

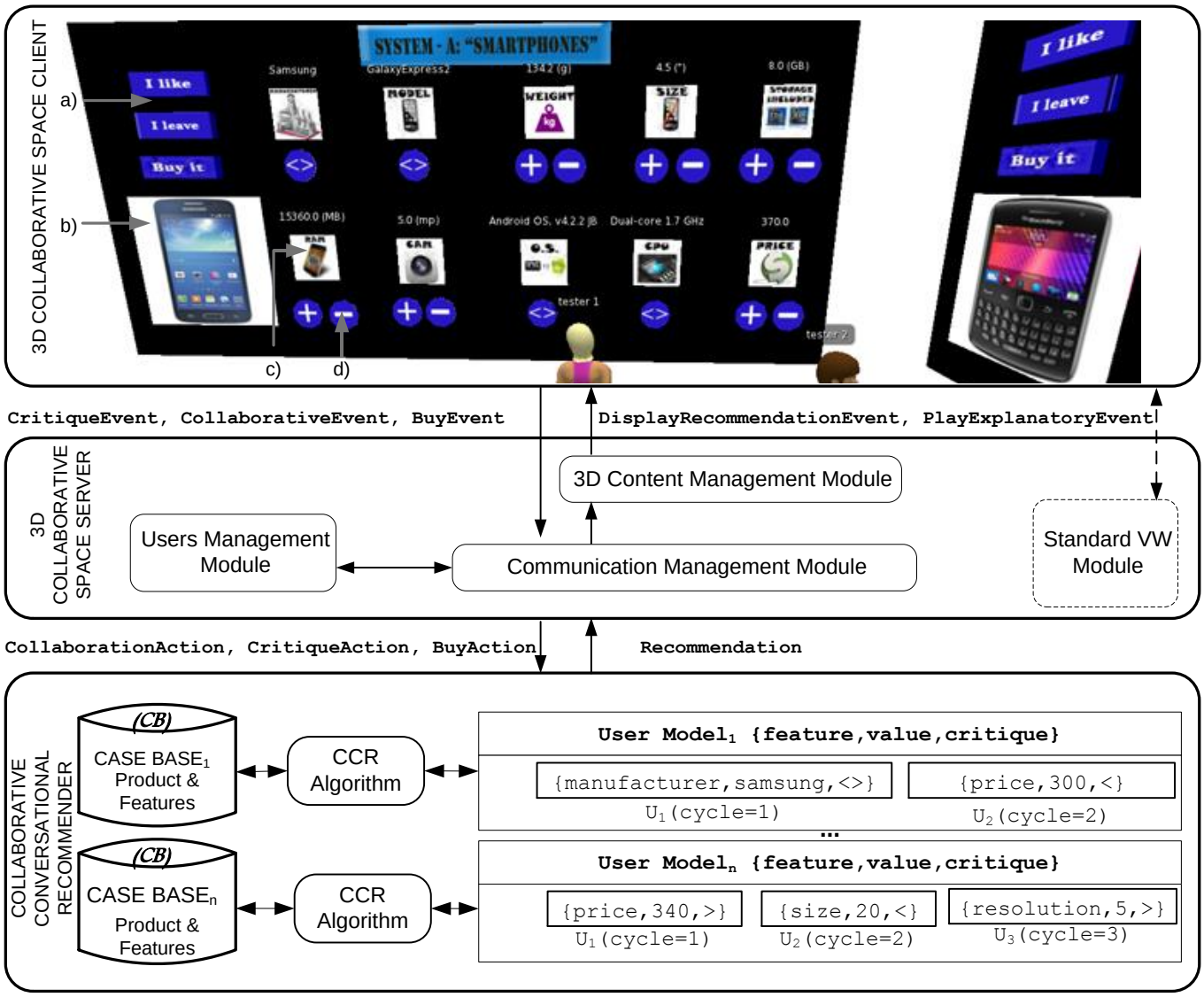


Fig. 1. 3 layers of the CCR framework

Space Client layer) with an initial product. This is done by means of the communication layer which connects both layers. product, for example price i.e. "I want a less expensive camera". She performs the critique by interacting with the visual elements in the interface, see the *3D Collaborative Space Client*. Next, the critique is sent as a message to the communication layer (*3D Collaborative Space Server*) that sends the critique to the *Collaborative Conversational Recommender (CCR)*. In this layer, the CCR Algorithm selects the next product recommendation based on the full set of products in the case base, *CB*, and the preferences stored in the user model, then, the recommendation is sent through the communication module to the user interaction layer (*3D Collaborative Space Client)* and the cycle starts again. The process finishes when the user either finds (and buys) the desired product or abandon the recommendation process.

These successive cycles represent a conversational recommendation process, where a sole individual interacts with the recommender. Nevertheless, the CCR framework contemplates a new recommendation process, described in depth in section 3.3, which supports both individual and collaborative critiques. Hence, users perform collaborative actions (*I like* and *I leave*) to start collaborating or stop the collaboration respectively. These actions provoke the transition between individual and collaborative critiquing states. On the one hand, a user can walk around the 3D space and see products recommended to other users. If she likes the product that the recommender is currently suggesting to another user, the *host user*, the *3D Collaborative Space Client* provides the former (*guest user*) with visual interactive elements to start the collaboration. Then, the *guest user* performs the so-called *I like* action in the *host user* recomender. After this action, both users change to collaborative critiquing state. In this state they talk each other

Then, the user can critique one feature of the

and arrive to a consensus for continue together the critiquing process. Nevertheless, to avoid that other users corrupt *host user* current product, she is in charge of directly interact with the recommender. On the other hand, at any time the *guest user* is free to continue alone the recommending process. This can be due to different reasons, for example, they do not arrive a consensus and the *guest user* prefers either to continue alone or to collaborate with other users. To this end, the user performs the so-named action *I leave* (in the *host user* recommender) and back to individual critiquing state. Note that a *host user* can be collaborating with any number of *guest users* at the same time.

In the following we detail each layer in the proposed CCR framework.

A. Collaborative Space Client

The *3D Collaborative Space Client* is an immersive 3D virtual space where users, represented as avatars, interact each other and with the recommender by means of a 3D Recommendation Object (RO). Therefore, this object facilitates user-recommender interaction.

The top of Figure 1 shows two users, each one situated in front of a 3D RO, which is a 3D panel consisting of several visual and interactive elements: a) corresponds to interactive elements for performing collaborative actions (*I like* and *I leave*) or for finishing (*Buy it*) the recommendation process when the user reaches the desired product, b) currently is an image of the recommended product, although it can be a 3D model visualization or a video of the current product, c) are visual affordances representing the features of the current recommended product, with the value of the feature on top of it, and d) displays one (<>, different than) or two (+,-) interactive icons the user touches for critiquing product features. For example, button <> is used to change nominal features like manufacturer and + and – buttons for critiquing numerical ones, i.e., "I want a different (<>) manufacturer", "I want a cheaper (-) camera" "I like a more expensive (+) camera".

User interactions in this layer trigger three types of events which are sent to the *Collaborative Space Server*: the `CritiqueEvent` when the user performs a *critique* by touching the visual element annotated as d) on the 3D RO; the `CollaborativeEvent` when the user wants to collaborate with another user and to do so she performs the *I like* action or, in contrary, she performs the *I leave* action to finish the collaboration; and finally, the `BuyEvent` when the user has found a suitable product, she performs the *Buy it* action.

This layer also receives two events from the *Collaborative Space Server*: the `DisplayRecommendationEvent` which is in charge of displaying a new product recommendation on the 3D RO, and the `PlayExplanatoryEvent` which informs the user that a suitable product has been found, through text chat and sound reproduction.

B. Collaborative Space Server

This layer fulfills the standard functions of a 3D virtual world server, which support the execution of an online multi-user 3D environment (see dotted squared on the right part of middle layer in Figure 1). Additionally, it incorporates three modules which facilitate the collaborative conversational recommendation: communication, users, and 3D content management. Next, we introduce them.

The *Communication Management Module* maps user events to recommender actions and, in reverse, recommender actions to user events. To do so, it requires user information from the *Users Management Module*, which stores and manages users' information such as user identification and state (individual or collaborative critiquing), and the 3D RO the user is interacting with.

In the following we present the three user events (introduced in section 3.1) which are mapped to recommender actions.

First, the `CritiqueEvent` maps to the `CritiqueAction` described in Equation 1. It contains: the user who performed the critique, *userId*, the recommender where was performed the critique, *recommenderId*, the current recommended product, *productId*, the critiqued feature, *featureId* (e.g. *price*), the type of critique, *typeCritique* (i.e.<>,+ , or -), and the critique value, *critiqueValue* (i.e. the current value of the critiqued feature).

$$CritiqueAction(userId, recommenderId, productId, featureId, typeCritique, critiqueValue) \quad (1)$$

For example, *critiqueAction(user2, rec2, camera 300, price, +, 340)* describes that *user2* working on *rec2* sends a critique about *product300* for obtaining a recommendation product with a price higher than *340*. Note that this critique is later stored in the user model, see Figure 1.

Second, the `CollaborativeEvent` maps to the `CollaborationAction` described in Equation 2, which contains the collaborative action, *actionId (I like* or *I leave)*, the user who performed the collaborative action, *userId*, the 3D RO where was performed the action, *recommenderId*, and the current recommended product, *productId*.

$$CollaborationAction(actionId, userId, recommenderId, productId) \quad (2)$$

Third, the `BuyEvent` maps to the `BuyAction` described in Equation 3, which involves the user who performed the action, *userId*, the 3D RO where was performed the action, *recommenderId*, and the product bought by the user, *productId*.

$$BuyAction(userId, recommenderId, productId) \quad (3)$$

Now we depict the `Recommendation` action which maps to the `DisplayRecommendationEvent` previously introduced in section 3.1. It contains – as shown in Equation 4 – the user who performed the critique or selected a product from other user, *userId*, the new recommended product, *productId*, and a list of features (i.e. the value for each product features), *featureValues*. Later, the mapped event is sent to the *3D Content Management Module*.

Recommendation(*userId*, *productId*, *featureValues*) (4)

The *3D Content Management Module* is in charge of displaying, `DisplayRecommendationEvent`, a new product recommendation on the 3D RO, and playing, `PlayExplanatoryEvent`, the sound which informs the user that she has found a suitable product.

C. Collaborative Conversational Recommender

CCR layer contains the following elements: recommendation algorithm (CCR algorithm), case bases (*CB*) and users models needed for the recommendation process. CCR algorithm is based on Incremental Critiquing (IC) [14] algorithm. However, the CCR algorithm includes critiquing and *selection* of product candidates from a collaborating user (called a *host user*). Particularly, in the CCR the set of products or cases for recommendation is defined as a case base $CB = \{p_1, ..., p_n\}$ where p_i is the ith product. Additionally, in the CCR the user model $U = \{U_1, ..., U_k\}$ contains a set of critiques where each U_j is the jth critique.

Figure 2 shows the CCR process in a cycle that maintains

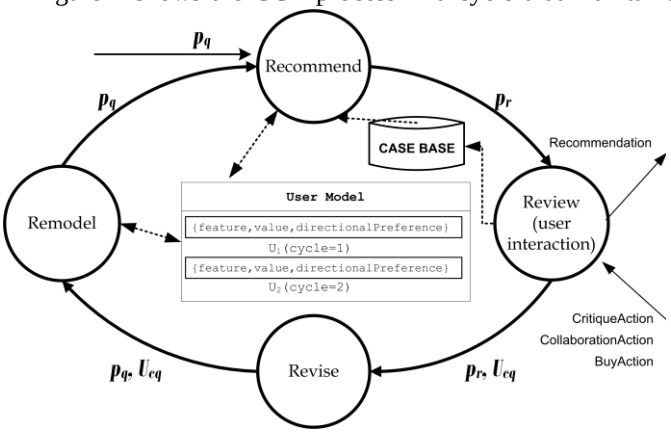


Fig. 2. CCR Cycle

four phases (i.e. recommend, review, revise and remodel) as IC but they differ in the internal process of these phases.

First phase is devoted to **recommend** a new product, p_r, to the user from the case base. This recommendation comes from one of the following options: (1) the recommended product is a selection made using a `CollaborationAction` (I like) when initiating a collaboration with a *host user*; or (2) the recommendation is based on current query, p_q, and previous *critiques* if there are any in the user model. In both options, this phase returns a product recommendation, p_r.

Specifically, in the last option we have kept from IC the idea that instead of ordering the relevant products[1] on the basis of their similarity to the product query (p_q) it is also helpful to compute a *compatibility* score, $C_{p_i}(U)$. We have maintained the $C_{p_i}(U)$ defined in IC but the CCR also includes compatibility scores based on reinforcement learning [16]. In particular, in this paper the compatibility score used is shown in the following equation:

$$C_{p_i}(U) = \frac{\sum_{j=1}^{|U|} \delta(p_i, U_j)}{|U|} \quad (5)$$

where p_i is the *ith* candidate product and $U = \{U_1, ..., U_k\}$ is a user model where each U_j is the jth critique and $|U|$ is the number of critiques in U. The satisfaction function δ returns 1 if case, p_i, satisfies critique U_j or 0 otherwise. Thus, the compatibility score is essentially the percentage of critiques in the user model that product p_i satisfies. Then, the compatibility score and the similarity of a candidate product, p_i, to current product query, p_q, are combined in order to obtain an overall *quality* score, Q:

$$Q(p_i, p_q, U) = \beta \cdot C_{p_i}(U) + (1-\beta) \cdot S(p_i, p_q) \quad (6)$$

where S is the similarity function based on an Euclidean distance, and β is set to 0.75 by default just as IC algorithm [14]. The quality score Q is used to rank the relevant products prior to next cycle, and the product with the highest quality is then chosen as the new recommendation, p_r.

In the **review** phase, a `Recommendation` is sent to the 3D Collaborative Space Server (see Figure 1) which maps to an event showing the recommended product in the 3D Collaborative Space Client and the user *reviews* current recommendation, p_r, by introducing some feedback. There are three feedback mechanisms available: the `CollaborationAction`, the `CritiqueAction`, and the `BuyAction`. The first feedback occurs with `CollaborationAction`, which generates a product selection for the user that starts a collaboration with another user, the *host user*. The second mechanism ensues when the `CritiqueAction` is performed through a critique element in the 3D interface (see 3D Collaborative Space Client in Figure 1). Anyone of them generates a new critique U_{cq} (i.e. a directional preference over a feature) to be considered in future recommendations. The last feedback is the `BuyAction`, which generates a product selection for the user and denotes that the CCR cycle will be finished as the user has found a suitable product. In addition to receiving feedback, this review phase is in charge of removing current product recommendation, p_r, from the case base, *CB*, for avoiding repetitions in subsequent recommendation cycles.

The third phase focuses on the **revision** of the current product query, p_q. Concretely, it defines the current product recommendation, p_r, as the new product query p_q.

Finally, the CCR cycle finishes **remodeling** the user model (U) according to the user's feedback provided in the review

[1] A relevant product is a product that satisfies the last critique made by the user.

phase. Maintaining a user model is not as simple as storing a list of previously selected critiques. Some critiques or the selected product may be inconsistent with earlier critiques. It is essential to remodel the user model by adding the latest critique (U_{cq}) if there is any only after pruning previous critiques so as to eliminate these inconsistencies. According to this need, this phase includes two steps.

The first step is devoted to remove those critiques that contradict current feedback. That is, it removes all existing critiques that are inconsistent with the new critique U_{cq} if there is any or removes those that are not satisfied by the selected product. For example, in a camera recommendation process, if the user model contains a critique $[manufacturer, \neq, Sony]$ and the product selected from *host user* is a Sony camera, critique $[manufacturer, \neq, Sony]$ is removed from the user model as it contradicts current product recommendation. The second step reforms those critiques for which the new critique or the product selection is a refinement. For example, a user model with the following critique $[price, \leq, \$1500]$ that receives a selection whose price is $\$1000$, the critique will be refined to $[price, \leq, \$1000]$.

The CCR terminates the recommendation process either when the user retrieves a suitable product or when she explicitly finishes it with the (`BuyAction`).

IV. USER EVALUATION

This section reports results of user tests performed to evaluate our proposal. Our main goal is to assess the usability of the CCR framework. To do so, we focus on different usability criteria such as learnability, effectiveness, efficiency, and satisfaction. Note that we evaluate task effectiveness by means of the decision accuracy, which measures how well the recommender supports the user in the finding of the desired product.

Additionally, we want to get feedback about the usefulness of the approach and users' willingness to use a 3D interface for collaborative recommendations in the future.

D. Methodology

The test followed the Summative evaluation method and mainly focused on gathering quantitative data [3]. We aimed to evaluate how the collaborative framework would facilitate the users' task of finding a suitable product. Then, we recruited 20 participants who performed the test in pairs.

In particular, the test protocol consisted of four phases: (1) A *pre-test interview* where the users were welcomed and introduced to the test. They also were asked about their experience with 3D virtual worlds and conversational recommender systems; (2) A *training* phase where users carried out a task that involved the execution of several actions within the virtual world, such as searching for a product using one recommender and making a collaboration with another user. This training phase was fully guided by the moderator; (3) A *test* phase where the users performed test tasks (described next in section 4.2) without receiving any guidance. Finally, a (4) *post-test questionnaire* phase where the moderator gave the users a questionnaire. This questionnaire consisted of ten questions (see Table I) and a free-text space for comments. The users answered these questions using a seven-point likert scale where 1 corresponds to "strongly disagree" and 7 to "strongly agree".

TABLE I
POST-TEST QUESTIONNAIRE

QUESTION NUMBER	STATEMENT
Q1 (Learnability)	It has been easy to learn how to interact with the recommender in the 3D virtual environment.
Q2 (Learnability)	It has been easy to learn how to interact with another user in the 3D virtual world interface.
Q3 (Collaboration)	I found searching for a product when I have been collaborating with another user more rewarding and entertaining.
Q4 (Perceived efficiency)	I feel that it took me less time to arrive to a target when I have collaborated with another user and I have selected their product.
Q5 (Usefulness)	I found the 3D interface to collaborate with other users useful.
Q6 (Perceived accuracy)	I believe that the product bought in each session was the best for me.
Q7 (Uselfulness)	I found the 3D collaborative interface useful for buying products using a recommender system.
Q8 (Effectiveness)	The recommender assisted me in the achievement of (buying) tasks, it had all the functions and capabilities I expected it to have.
Q9 (Satisfaction)	Overall, the system was easy to use.
Q10 (Intention to use in the future)	I will use this system for buying products in the future.

The evaluation was performed using a SMARTPHONE data set, which consists of 1722 SMARTPHONE products with 5 nominal and 9 numerical features (i.e. manufacturer, model, length, width, profundity, weight, size, multi-touch capability, storage capability, ram, resolution, operating system, cpu, and price).

E. User Tasks

Users were requested to perform the three following tasks:

Task 1. A collaborative recommendation task with similar target products. This task was defined in such a way that the targets (i.e. the product the user is searching for) were predefined for each user and these targets shared some similarities. However, users were unaware of that fact.

Task 2. A collaborative recommendation task with dissimilar target products. This task is equivalent to the previous task with the particularity that this time targets were unalike. Again, users were ignorant of that fact.

Task 3. A freely collaborative recommendation task, where a target product was not predefined and users could freely decide to buy a desired product. The goal of this task is to measure the decision accuracy of our proposal, which is related to the effectiveness usability criterion. Therefore, after the user has found and bought the desired product using CCR framework, she reviewed a full set of products and decided to select one of them, which could be the same product bought using the recommender or any other. In fact, we selected a subset (90 products) of the SMARTPHONE data set for facilitating the review to the user. As a result, decision

accuracy is the percentage of times that the user chooses to stick with the product bought using the recommender.

F. Results

First of all, we analyze post-test questionnaire results. Next, we report results on efficiency and decision accuracy of the proposal. Related to efficiency, we are interested in both user perceived efficiency and the actual efficiency, i.e. the number of cycles, or critiques from the user's point of view, they need to reach the desired product. The same interest applies to decision accuracy.

Figure 3 depicts the results obtained from the post-test

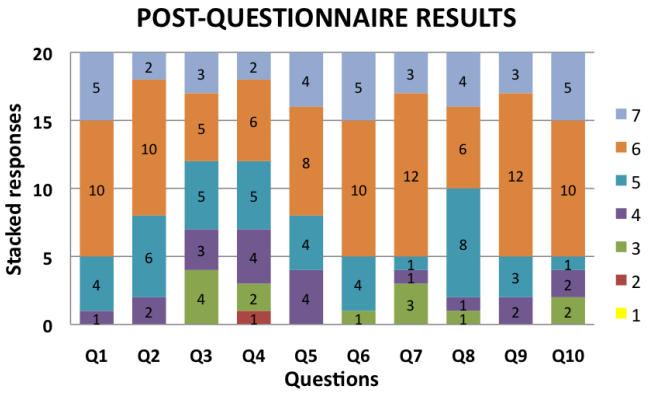


Fig. 3. Grouping rating for each one of the questions in the post-questionnaire

questionnaire (see Table I). Figure 3 is described as a stacked column chart that details for each question the number of responses received in a seven-point likert scale. Note that these results are related to the subjective perception of users but are quantitative data which give us valuable information about users' perception of usefulness and usability of our CCR framework.

Overall, the quantitative results obtained from the questionnaire were very satisfactory. It is worth noting that 83% of the responses were ranked with 5 or more points, none of the participants replied questions with a minimal score (1 value).

Considering the learnability of the CCR environment (i.e, questions Q1-Q2), participants' responses show that the users found the system easy to learn. Nearly all participants (19 over 20) ranked over 5 points Q1, which represents a 95% of the users. Moreover, in Q2 the evaluation is very satisfying too as 18 participants ranked over 5 points.

With regard to users' opinion about the ease of use of the recommender, results of Q9 show that 18 participants positively evaluated this aspect with more than 5 points. Furthermore, answers to questions Q7 and Q5 denote that users perceived that the recommender aided them in the searching of a product. Note that 16 participants ranked questions Q7 and Q5 with 5 or more points (75% of the participants).

In addition, user's opinion about the functions and capabilities of the system for aiding them to buy a product is very satisfactory (Q8 shows that only 2 participants ranked it with a value lower than 4 points). Moreover, when they were asked about their intention to use this system for a similar task in the future (Q10), 16 participants ranked this question with more than 5 points, which means that users have a good perception about the usefulness of such an integration of a collaborative conversational recommender (CCR) within a 3D virtual world. Moreover, during the test, users were comfortable when collaborating with other users (Q3 with 13 participants over 5 points).

Related to users' efficiency perception during the test, users felt they finished the recommendation process in less time when they collaborated with other users. Thus, the majority ranked with 5 o more points question Q4 (13 participants). These results corroborate those of task efficiency in terms of average session length (ASL or number of cycles), as described next.

Figure 4 shows efficiency data gathered in the three tasks previously introduced in section 4.2. Note that users started

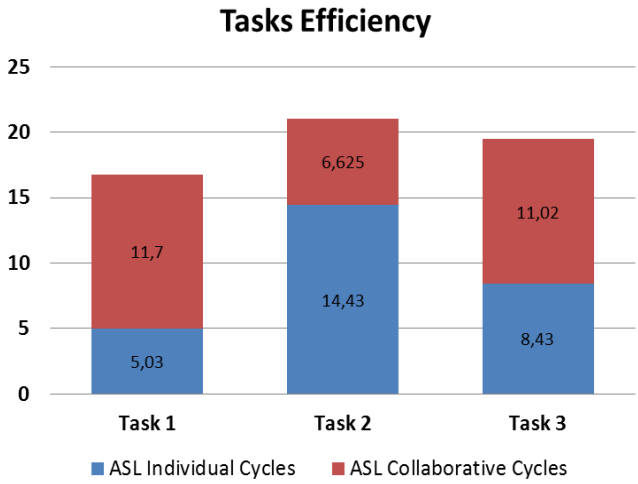


Fig. 4. Average recommendation efficiency for the three tasks of the experiments

the recommendation process individually and then suggested to collaborate with the other user (if they wanted). Then, Figure 4 depicts both individual and collaborative cycles for each task.

A *collaborative recommendation* with similar preferences in the target (Task 1) obtains the lowest value in total number of cycles (16.74 cycles), whereas the task with dissimilar preferences (Task 2) in their targets enlarges the ASL until 21.05 cycles and the task without target (Task 3) reaches 19.45 cycles. For Task 1 and Task 2, this is expected as the collaboration (ASL Collaborative Cycles in red) is reduced from 11.70 cycles in Task 1 to 6.63 cycles for the Task 2, probably due to the users realised that targets were dissimilar and so preferred to continue more time interacting individually. The number of collaborative cycles and the number of individual cycles were more balanced in Task 3, likely users started individually with an idea more or less clear of their desired products but some time after realised this idea was not clear enough and then decided to collaborate.

User's perceived accuracy is reported in question Q6 whose results denote that 15 of the 20 users rated it, within a range

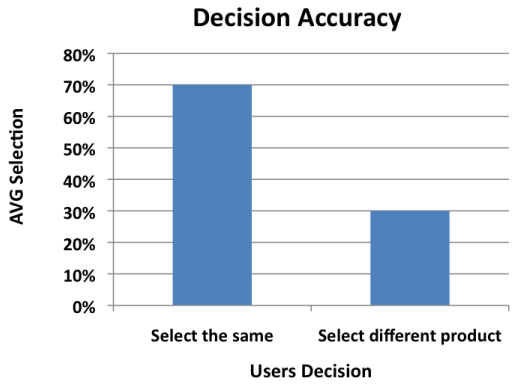


Fig. 5. Average decision accuracy

between 6 and 7, i.e. their final product selected using the CCR framework is the best for them. Put it differently, 75% of users perceive that their final selection has been accurate enough. If we include those that consider this question with a 5 value in the scale, the satisfied users with their final selection increase until 95%.

Additionally, we have also measured the decision accuracy with Task 3, as it has been previously described in Section 4.2. Figure 5 shows a relative higher *decision accuracy* measure in the CCR algorithm, which achieved 70%. This measure means that 70% of users bought the same product using the recommender and later using the full set of products. The remaining 30% of users switched to a different, better choice when they had the opportunity to view all the products. These results show the effectiveness of our approach, which aids users in finding their desired products.

V. CONCLUSIONS

In this paper, we have proposed a new framework to integrate a Collaborative Conversational Recommender (CCR) in a 3D social virtual world. Additionally, we have described in depth the three layers of the CCR framework: the user interaction layer, the communication layer, and the recommendation layer. We carried out a user evaluation of our proposal with three different tasks. The results are positive attending to different usability criteria, such as learnability, efficiency, effectiveness perceived and satisfaction. In the post-test questionnaire 83% of the responses were ranked with 5 or more points and none of participants replied questions with a minimal score. Additionally, efficiency measures revealed that users benefited more from the collaborative interaction when they had targets with similar preferences. Effectiveness, or decision accuracy results showed that the CCR framework aids users in finding their desired products. To date our proposal is the first approach that integrates a content-based recommender algorithm within a 3D virtual environment.

As future work we plan to incorporate a new user-recommender interaction style based on natural language. We also aim to incorporate, when possible, data sets with 3D models of products.

ACKNOWLEDGMENT

D. Contreras has been supported by a doctoral fellowship "Becas Chile" from CONICYT of the Chilean Government. This research has also received support from the project TIN2012-38603-C02, CSD2007-0022, TIN2011-24220 and TIN2012-38876-C02-02 from the Spanish Ministry of Science and Innovation.

REFERENCES

[1] D. Aha, L. Breslow, and H. Muñoz Avila. Conversational Case-Based Reasoning. *Applied Intelligence*, 14:9–32, 2001.

[2] A. Arroyo, F. Serradilla, and O. Calvo. Modeling spatial-temporal context information in virtual worlds. In *Proceeding of the 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems*, pages 437–447. Springer, 2010.

[3] D. Bowman, J. Gabbard, and D. Hix. A Survey of Usability Evaluation in Virtual Environments : Classification and Comparison of Methods. *Presence: Teleoperators and Virtual Environments*, 11(4):404–424, 2002.

[4] R. Burke. A Case-Based Reasoning Approach to Collaborative Filtering. In E. Blanzieri and L. Portinale, editors, *Proceedings of the Fifth European Conference on Case-Based Reasoning, EWCBR '00*, pages 370–379. Springer, 2000. Trento, Italy.

[5] R. Burke. Interactive critiquing for catalog navigation in e-commerce. *Artificial Intelligence Review*, pages 1–21, 2002.

[6] D. Contreras, M Salamó, I. Rodríguez, and A. Puig. Integrating a Conversational Recommender System within a 3D Collaborative Environment. In *Proceedings of the XV Conference of Spanish Association for Artificial Intelligence*, pages 89–98, 2013.

[7] G. Guo and M. Elgendi. A New Recommender System for 3D E-Commerce: An EEG Based Approach. *Journal of Advanced Management Science*, 1(1), 2013.

[8] M. Holzwarth, C. Janiszewski, and M. M Neumann. The Influence of Avatars on Online. 70(October):19–36, 2006.

[9] X. Hu and K. Wang. Personalized Recommendation for Virtual Reality. *2010 International Conference on Multimedia Technology*, pages 1–5, October 2010.

[10] K. McCarthy, M. Salamó, L. Coyle, L. McGinty, B. Smyth, and P. Nixon. Group recommender systems: A critiquing based approach. In *Proceeding of the 11th International Conference on Intelligent User Interfaces*, pages 267–269. ACM Press, 2006.

[11] Lau, H. F., Kan, C. W., & Lau, K. W.. How consumers shop in virtual reality? How it works? Advances in Economics and Business, 1(1), 28-38, 2013.

[12] P. R. Messinger, E. Stroulia, K. Lyons, M. Bone, R. H. Niu, K. Smirnov, and S. Perelgut. Virtual worlds at past, present, and future: New directions in social computing. *Decision Support Systems*, 47(3):204 – 228, 2009. Online Communities and Social Network.

[13] P. Pu, L. Chen, and P. Kumar. Evaluating product search and recommender systems for e-commerce environments. *Electronic Commerce Research*, 8(1-2):1–27, 2008.

[14] J. Reilly, K. McCarthy, L. McGinty, and B. Smyth. Incremental critiquing. *Knowledge-Based Systems*, 18(4-5):143–151, 2005.

[15] F. Ricci and Q. Nguyen. Acquiring and Revising Preferences in a Critique-Based Mobile Recommender System. *IEEE Intelligent Systems*, 22(3):22–29, 2007.

[16] M. Salamó and S. Escalera. Increasing retrieval quality in conversational recommenders. *IEEE Transactions on Knowledge and Data Engineering*, 24(10):1–14, 2012.

[17] F. Shah, P. Bell, and G. Sukthankar. A destination recommendation system for virtual worlds. In *FLAIRS Conference*, pages 475–476. AAAI Press, 2010.

[18] B. Xu and Y. Yu. A Personalized Assistant in 3D Virtual Shopping Environment. *Second International Conference on Intelligent Human-Machine Systems and Cybernetics*, pages 266–269, 2010.

[19] J. Zhang, N. Jones, and P. Pu. A Visual Interface for Critiquing-based Recommender Systems. *Proceeding of the 9th ACM Conference on Electronic commerce*, pages 230–239, 2008.

[20] O. Sanjuán, B. C. Pelayo G-Bustelo, R. G. Crespo, E. Torres-Franco. Using Recommendation System for E-learning Environments at degree level. *International Journal of Interactive Multimedia and Artificial Intelligence*,1(2):67-70, 2009

David Contreras received his MTech CSE from the University Federico Santa María, Valparaíso, Chile. He received an award for his Ph.D. from Doctoral Fellowship "Becas Chile" of the Chilean Government. He is currently an Assistant Professor in Faculty of Engineering and Architecture of the University Arturo Prat, Iquique, Chile. His research interests include recommender systems, virtual environments, and user modeling.

Maria Salamó received both her B.S. in Computer Science (1999) and her Ph.D. (2004) degrees from the Universitat Ramon LLull, Barcelona, Spain. She received an award for her Ph.D. project on Information and Communication technologies. She is currently a lecturer at the Department of Applied Mathematics and Analysis, Universitat de Barcelona. Her research covers a broad range of topics within AI including recommender systems, case-based reasoning, user modeling and personalization techniques.

Inmaculada Rodriguez Santiago studied computer science in the University of Granada (Spain) and obtained her Ph.D in the University of Alcala (Spain) in 2004. She is associated professor in the Applied Mathematics Department in the University of Barcelona where she is member of WAI (Volume Visualization and Artificial Intelligence) research group. She has worked in the computer animation of virtual humans and currently her major research focus lies on the integration of both artificial intelligence and virtual worlds technologies, serious games and gamification.

Anna Puig received her BSc. in Computer Science (1991) and her PhD in Computer Science (1998) from he Polytechnic University of Catalonia. Currently, she is associated professor in the University of Barcelona (Spain) . She researches in medical volume visualization as well as the usage of Artificial Intelligence methods in Virtual Worlds. She has experience in research projects about medical visualization, game-based learning and the popularization of mathematical activities addressed to primary and secondary schools.

Using Local Grammar for Entity Extraction from Clinical Reports

Aicha Ghoulam[1], Fatiha Barigou[2], Ghalem Belalem [3], Farid Meziane[4]

[1, 2,3]*Departmnt of Computer Science, Faculty of Exact and Applied Sciences, University of Oran, Ahmed BenBella, Algeria*
[4]*School of Computing, Science and Engineering, University of Salford, Manchester, UK*

***Abstract* — Information Extraction (IE) is a natural language processing (NLP) task whose aim is to analyze texts written in natural language to extract structured and useful information such as named entities and semantic relations linking these entities. Information extraction is an important task for many applications such as bio-medical literature mining, customer care, community websites, and personal information management. The increasing information available in patient clinical reports is difficult to access. As it is often in an unstructured text form, doctors need tools to enable them access to this information and the ability to search it. Hence, a system for extracting this information in a structured form can benefits healthcare professionals. The work presented in this paper uses a local grammar approach to extract medical named entities from French patient clinical reports. Experimental results show that the proposed approach achieved an F-Measure of 90. 06%.**

***Keywords* — Information Extraction, Electronic clinical reports, medical entities recognition, natural language processing.**

I. INTRODUCTION

RECOGNITION and classification of named entities in texts is recently considered as an important task in automatic natural language processing (NLP) as they play a significant role in various types of NLP applications, especially in Information Extraction, Information Retrieval, Machine Translation, Question-Answering and , Parsing/Chunking.

The amount of information written in natural language and available in electronic forms is increasing, making the development of intelligent tools to process this information an urgent need for practitioners such as health care professionals. Information Extraction is gaining an increased attention by researchers, who seek to acquire knowledge from this huge amount of natural language content. Many approaches have been proposed to extract valuable information from texts in different fields, with the medical domain being one of them.

We note that the volume of medical information is constantly increasing. According to [1] it doubles every five years and this wealth of information is difficult to access because it is stored in unstructured formats. This is particularly true for the case of clinical reports (CRs) where information such as pathologies, medical history and diagnoses are recorded in a textual format, is ever increasing and becoming difficult to search and access. CRs can have a positive impact on the quality of care, patient safety and efficiencies in medical procedures. However, without accurate and appropriate content in a usable and accessible form, these benefits may not be achieved. Developing a system for extracting unstructured information can benefits healthcare professionals.

These kinds of systems have become very necessary tools; they will enable researchers to access accurate data and the required information, and reducing the time spent by doctors on making decisions about patients' diseases. Hence, the main motivation of this work is to develop an automated system for extracting named entities from clinical reports.

Firstly, most of the elements in CRs are name entities (e.g.., the names of patients, diseases, symptoms, and drugs) that can be used in various applications, such as seeking information to diagnose new patients, conducting epidemiological studies, statistical analysis, and data mining. However, these CR are difficult to analyze due to their unstructured nature and the large volume of records available. Secondly, most existing medical entities extraction systems are devoted to English. Research in the French language is still at its initial stages [2].

In this research, we propose to use an original approach based on local grammars to extract medical entities from French clinical reports. The local grammar based approach has recently been applied to extract proper nouns in English, Chinese, French, Korean, Portuguese, and Turkish news texts [3]. This approach was first used to discuss recursive phrases that are commonly found in specialist literature like biochemistry and then extended to extract date, time and address expressions from letters.

In this work, we study the application of local grammars for extracting medical entities from French clinical reports. We focus on the extraction of the following named entities; disease, symptom, treatment, drugs and clinical reviews. The rest of the paper is organized as follows Section 2 summarizes the task of named entity extraction and work related to the medical field. In section 3 we describe information extraction and methods. Section 4 describes the proposed system and our contribution to extract medical entities. Section 5, presents evaluation results concerning proposed system performance.

DOI: 10.9781/ijimai.2015.332

Finally, this paper ends with conclusions and some ideas for future works.

II. RELATED WORKS

The Named Entity Extraction (NEE) task aims to recognize named entities and classify them into categories like organization names, person names, location names, date and time expressions, monetary amounts and documents' references [4]. Named Entity Extraction systems are based on two main approaches: the rule based and the machine learning approaches [5]. Hybrid systems combine these two approaches [6].

The Rule-Based approach is a manual technique based on a specific domain extraction rules written by an expert using morphological and syntactic information like trigger words, capitalization, and gazetteer. This approach gives very good results however requires great human efforts and a considerable time for data analysis and rule writing. It is time consuming during development and it lacks portability which limits its extension to other domains.

On the other hand, the machine learning approach, is a trainable technique that is capable to improve its ability to extract information from input automatically or under supervision, but it requires large annotated corpora for training, which are both expensive and time-consuming to train the models [7]. Many different models have been proposed over the years. The most prominent of these are Hidden Markov Models (HMM) [8], Support Vector Machine (SVM) and Conditional Random Fields (CRF) [9].

Several studies have used the NEE task, [5,7,10]. Most systems were mostly interested with named entity like organization names, place names, date expressions and numeric expressions [11] with different languages [12] and gave promising result. Recently, NEE has been applied to the medical field to extract entities such as protein names, gene names, disease names and treatments from medical documents [7]. Various systems have been developed, using rule-based approaches, including MedLEE [13], SymTex [14], MetaMap [15] and MedIX [16].

The MedIX system [16] was applied to patient CRs using natural language processing techniques. It performs some processes such as preprocessing the text, tokenizing, and tagging, recognizing special formatting and then it identifies entities and classifies them into categories that included patient name, disease name and symptom names. Others classify entities into problem, treatment, test classes [9] and drug properties [17].

Authors in [18] proposed an approach relying on linguistic pattern and canonical entities to extract five categories of medical entities from CRs namely, disease name, treatment name, drug name, and test and symptom names. Other systems extract useful entities from radiology and mammography reports to identify patients with lung cancer [19] or with tuberculosis [20].

Recent studies are mostly based on the machine learning approach; [1] and [21] employ support vector machines to attribute semantic categories to each word in discharge summaries. Markov models-based methods are also frequently used [8]. Others used unsupervised methods were based on seed term collection [22].

In the past couple of years, researchers have been exploring the use of machine learning algorithms in the clinical concept detection. To promote the research in this field many organizations such as ShARe/CLEF, SemEval have organized a few clinical NLP challenges. Both rule based [23,24,25] and machine learning based methods as well as hybrid methods [26,27,28,29] were developed. In this shared-task sequential labeling algorithms (i.e., CRF) [30,31,32,33,34,35], and machine learning methods (i.e., SVM) [36] have been demonstrated to achieve promising performance when provided with a large annotated corpus for training.

The system that was top-ranked in the SemEval 2014 Task 7 among all participating teams is given in [32]; authors developed three ensemble learning approaches for recognizing disorder entities consisting of an ensemble learning-based approach and a Vector Space Model based method for disorder entity encoding. Extracted features from clinical notes were used to train two machine learning algorithm-based entity recognition models, CRF and Structural Support Vector Machines (SSVMs). These two models were ensembled with MetaMap. Their approaches achieved top rank in both subtasks (disorder entities recognition and encoding), with the best F-measure of 81.3% for entity recognition and the best accuracy of 74.1% for encoding, indicating that their proposed approaches are promising.

Another work [37] presented a comparison of two approaches to automatically de-identify medical records written in French; rule based system and CRF based system. They achieved an F-measure of 84.3% and 88.3% for each system respectively in cardiology reports. They achieve an F-measure of 68.1% and 63.8% for each system respectively in fœtopathologie reports. They concluded that a rule based system allowed them to achieve good results on nominative and numerical data, and the machine learning approach performed best on more complex categories.

III. INFORMATION EXTRACTION AND METHODS

Information Extraction (IE) has been defined in the literature by many researchers [38, 39]. The most common definition is that IE is an automatic process for extracting structured information which can be relevant for a particular domain from unstructured documents like free text that are written in natural language (e.g. news article, clinical reports) or semi-structured documents that are pervasive on the Web, such as tables or itemized and enumerated lists. The obtained data are then arranged to be incorporated into machine readable databases and ontologies which, in turn, are used to improve applications such as Question Answering engines or Information Retrieval systems.

Five separate component tasks, which illustrate the main functional capabilities of current IE systems, were specified by recent MUC-7 evaluation [5]:

- Named Entity Recognition (NER), involves the recognition of named entities such as organizations, persons, locations, dates and monetary amounts. In the clinical domain, this might include entities such as disease and drug.
- Relation Extraction (RE) task; is the task of detecting and characterizing the semantic relations between entities in text. In the clinical field, it includes for example relation between disease and drug.
- Co-reference Analysis task, is a task which determine linguistic expressions that refer to the same real-world entity in natural language, has not yet been widely applied to clinical documents [40].
- Template Filling, the information to be extracted like entities, relationships and events in natural language texts is pre-specified in user defined structures called templates (or objects), each consisting of a number of slots (or attributes), which are to be instantiated by an IE system as it processes the text.
- Event Description, [41] defined a medical event as anything that is clinically important and that can also be mapped to a timeline. They created the i2b2 2012 challenge; a clinical temporal relation corpus that includes clinical events, temporal expressions, and temporal relations.

An information extraction system supports one of the two basic methods of extraction, namely, rule-based information extraction method, and statistical information extraction method.

- The Rule-Based IE methods: rule-based methods extract the information by rules, and these rules can be generated by human hand-coded, or by learning from examples. The most representative examples of this kind of systems are FASTUS [42], GE NLTOOLSET [43], PLUM [44] and PROTEUS [45]; these systems are well described in [46]. They can achieve good performance on the specific target domain. Human hand-coded rule-based system, in some sources also called knowledge engineering method, gives very good results. However, it involves a great human effort and a considerable time for data analysis and rule writing. It is time consuming during development [55].
- Statistical learning IE methods: statistical learning methods or Machine Learning (ML) methods; are trainable techniques able to improve their ability to extract information from input automatically or under supervision see the survey of [5]. Most recent studies use supervised machine learning starting from a collection of training examples; the idea of supervised learning is to study the features of positive and negative examples of information to be extracted (e.g. entities, relations, attributes) over a large collection of annotated documents and design rules that capture instances of a given type. Many different models have been proposed over the years. The most prominent of these are (HMM), Maximum Entropy Markov Models (MEMM), SVM or even a vector classification model for which the features are not terms, but graph metrics [47] and CRF. Other studies used unsupervised machine learning methods; a class of problems in which one seeks to determine how the data are organized such as clustering; a common technique for statistical data analysis used in many fields as used in [48].
- Wrapper induction: many approaches for data extraction from web pages have been developed to transform the web pages into program-friendly structures such as a relational database. Wrapper induction system considers web pages as a source data. It is a program that wraps an information source like a database server, or a web server [49]; it usually performs a pattern matching procedure like a form of finite-state machines which relies on a set of extraction rules.
- IE using Ontology: Ontology is a formal and explicit specification of a shared conceptualization; it plays a crucial role in the process of IE. The relation between ontologies and IE is involved in two tasks: on the one hand, Ontology is used for information extraction; IE needs ontologies as part of the understanding process for extracting relevant information [50]; on the other hand, information extraction is used for populating and enhancing a domain ontology from the web as shown in [51]; they developed an ontology of a scene from the essential semantic components for the semantic structuring of the Web3D. The construction of ontology for the definition of tridimensional spaces will allow the Web3d to standardize the development of scenarios and the creation of manufacture agents that will make easier the modeling and texturing processes.

IV. PROPOSED APPROACH

In this study; we use and evaluate a rule based approach relying on local grammar the motivation and the description of this approach is presented in this section.

A. Benefits of the proposed system

Fig.1 show some benefits of such system for clinical staff. An UML use case diagram is used to describe the expected functionalities of the proposed system. Medical named entities recognition, as shown in Fig.1, is essential to built new systems to help doctor and clinical staff in their work. Doctors need quick and easy access to quality information resources to be able to make informed decisions regarding patient care; they also need systems to help them answer clinical questions.

1) Question-Answering systems:

i. Clinical staff asks to obtain medical response.
ii. A research in medical ontology must be done.
iii. The construction of medical ontology based on medical entity recognition and relation extraction between medical entities.

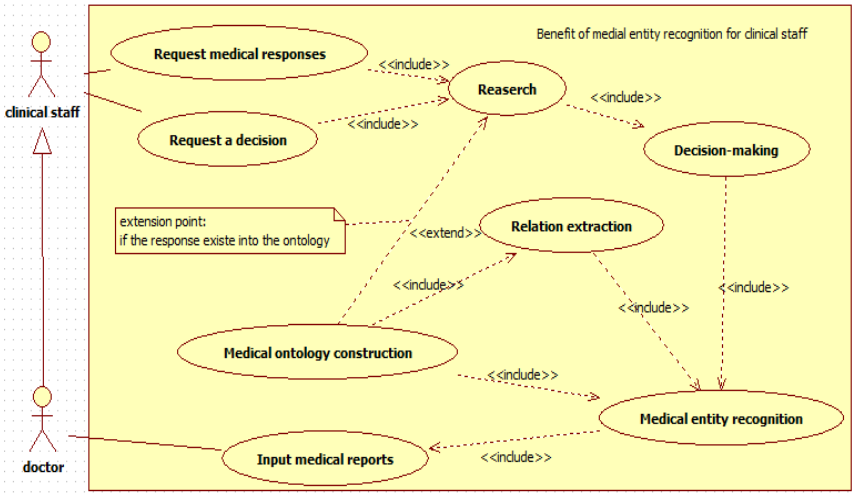


Fig.1. Medical entity recognition's benefits

iv. The extraction of relation between medical entities task necessities that the medical entities must be chucked.

2) Decision support system:

i. Clinical staff requests a decision.
ii. A research in past problems is done to make decision.
iii. Past problems input by doctors must be checked using medical entity recognition to facilitate research for similar cases.

B. Local grammar based approach

The Local Grammar (LG) approach was initiated by Harris [52] to discuss recursive phrases that are commonly found in specialist literature like biochemistry (immunology) [53]. Harris defines a local grammar as a way of describing syntactic restrictions of certain subsets of sentences which are closed under some or all of the operations in the language.

More specifically, LG is a way of recognizing the behavior of words that are used in a specific domain, finding how these words are used in sentences and inferring their usage patterns.

For example, Traboulsi [53] considered frozen expression as a subset of sentences that have some syntactic restrictions.

Certain expressions such as 'compound words' (e.g. stock market) are strictly frozen and others are partially frozen and are included in expressions such as the director of a small company, the director of a doctoral thesis as illustrated in the following patterns:

(financial + stock + E) market
Director of (company + thesis)
The 20 March (next + 2006)

Local grammar were extended by Gross [54] to extract date, time and address from letters. Gross defined LG as a finite state grammar and used it for finding words related by prefixation, suffixation, and sentences having similar syntax.

For certain expressions such as dates, times, and other types of proper names, it appears impossible to individually identify the set of all possible constructions and much more effective a

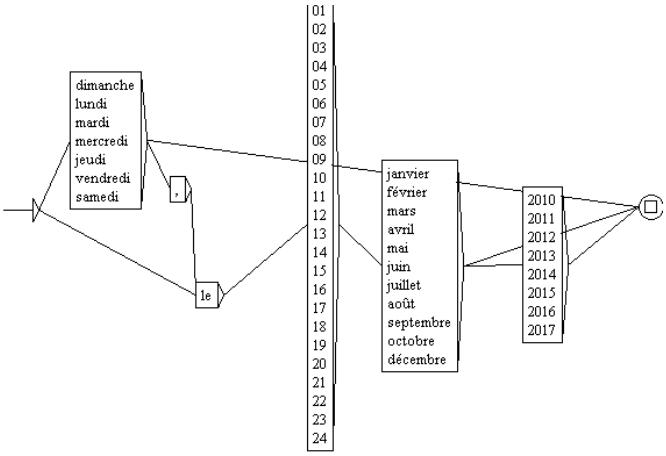


Fig.2. Example of a local grammar for French date expressions

representation in the form of automata. This representation is easy to be read of course if the graphs are well arranged. We give in Fig.2 an example of a local grammar for French date expressions.

It can recognize expressions like: "*dimanche, le 02 septembre 2014*".

Local grammars as finite state local automata have been used by [3] to recognize English person names in textual documents and then extended it to extract Arabic person names in [24].

C. Local Grammar based Approach for Extracting Named Medical Entities

In this work we study French CR to extract medical named entities using local grammar. In table 1, we gave the classes of entities and examples for each one.

We noticed that medical entities occur frequently at constructions having consistent structures in the proximity of Reporting Words (RWs) like "*consulte pour" (*consulting for), "*présentant*" (having) in the case of disease entities, "*signe de*" (sign of) in the case of symptom entities which are sufficiently frozen to be described in the form of local grammars. An example of these local grammars is shown in Fig. 3.

TABLE I
MEDICAL NAMED ENTITIES EXAMPLE

ENTITY	MEDICAL ENTITY EXAMPLE
Disease	Masse du pancreas (Mass of the pancreas)
Symptom	Anorexie (anorexia) Amaigrissement (weight loss) Déshydratation (dehydration)
Clinical Review	Scanner AP échographie Abdomino-Pelvienne (abdomino- pelvic ultrasound)
Treatment	Alimentation orale légère (Lightweight oral feeding) réhydratation 1 fl (rehydration 1 bottle)
Medication	Cefacidal 1g , Gentamicine 80 mg, Flagyl 1 fl

This graph is able to recognize constructions like:

- [Un malade nommée X présente une **fistule de fémur droite**]
  (A *patient named X has a* ***right femoral fistula)***
- [*Un malade Y consulte pour un* ***traumatisme lombaire***]
  (*A patient named Y consults for* ***lumbar trauma***)

The boxes labeled <disease>, <organ>, <location>, <adjective> are the names of sub-graphs that recognize candidates of disease names, organ names (anatomy), location, and adjectives respectively. Local grammar graphs containing sub-graphs shows similarity to recursive transition networks.

To extract medical entities from French clinical reports written in a free and natural language, our contribution adopts the following approach:

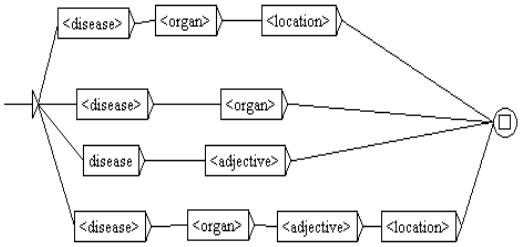


Fig.3. Example of the local grammars of disease entity

- Construction of different Gazetteers;
- Construction of medical entities classification rules;
- Describing the rules in the form of local grammars.

D. System Architecture

Figure 4 shows the architecture of the system. Our system has two major components: the gazetteers and the Grammars.

Pre-processing task:

It is necessary to properly delimit the clinical report into meaningful units. Most natural language processing solutions expect their input to be segmented into sentences, and each

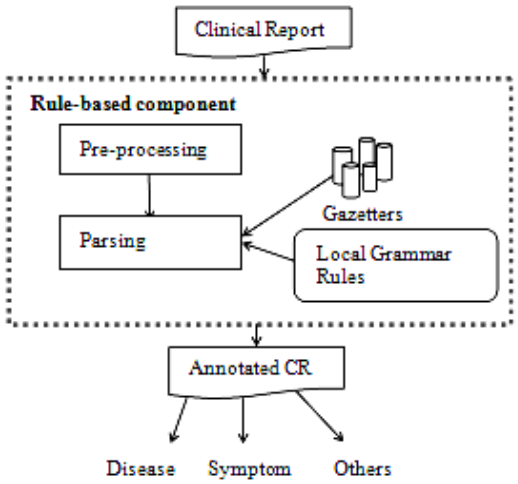


Fig.4. Architecture of the system

sentence into tokens; so for that we used the Unitex[1] open source for splitting CR into sentences and splitting sentences into tokens.

Gazetteers:

The gazetteer contains diseases names, symptoms names, clinical reviews, treatment and medications, medical adjectives, organs and so on. These Dictionaries are in electronic format; we have assembled them from different web site:

- A dictionary of adjectives[2] containing 514 entries.
- A dictionary of organ (Atlas: human body)[3] containing 384 entries.
- A dictionary of diseases[4,5] containing 343 entries.
- A dictionary of treatments[6].
- A dictionary of clinical reviews[6] containing 28 entries
- A dictionnaire of symptoms[7] containing 67 entries
- A dictionary of drugs[8,9]
- A list of French medical reporting words or trigger words
- A dictionary of medical names.

Grammars:

The grammar performs recognition and extraction of medical entities from clinical reports based on combination of regular expression patterns in the form of local grammars. A deep contextual analysis of various French clinical reports was performed using the Unitex open source software to build local grammars based on keywords or trigger words forming a window around medical entities.

1 http://www -igm.univ-mlv.fr/~unitex
2 http://www.linternaute.com/dictionnaire/fr/definition/abdominal/
3 http://www.doctissimo.fr/html/sante/atlas/index.htm
4 http://www.passeportsante.net/problemes-et-maladies-p69
5 http://www.vulgaris-medical.com
6 http://www.e-sante.fr/
7 http://www.vulgaris-medical.com/symptomes
8 http://www.eurekasante.fr/medicaments/alphabetique
9 http://www.doctissimo.fr/html/medicaments/medicaments.htm

TABLE IV
DETAILED EVALUATION ON THE CLINICAL REPORTS.
PRECISION (P), RECALL (R) AND F-MEASURE (F)

CATEGORY	P	R	F
Disease	0,921	0,800	0,856
Symptom	0,971	0,917	0,943
Treatment	1,000	0,765	0,867
Clinical Review	1,000	0,941	0,969
Drug	1,000	0,765	0,867

Example rule:

The following rule recognizes a disease name composed of medical name followed by a medical adjective and human organ based on a proceeding disease indicator pattern which is the RW.

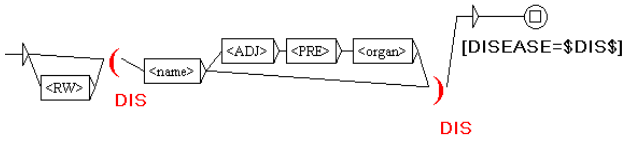


Fig.5. Example of Local Grammar in Unitex

(*name* + *ws*+ *adjectives* +*French prepositions* + *ws* + *organ(s)*)

Corresponding Local Grammar:

The following local grammar corresponds to the above rule using the Unitex editor:

Writing conventions:

- ws: whitespace.
- Name: dictionary of medical names.
- ADJ: dictionary of medical adjectives.
- PRE: dictionary of French prepositions.
- Organ: dictionary of human organ.

Example:

The following disease name would be recognized by the above local grammar shown in Fig.5:

"Masse tumoral du colon."; [Tumor mass of the colon]

We created a set of rules using Unitex to classify different medical named entities into disease, symptoms, clinical review, drugs and treatment from French clinical reports. Some examples of rules for each class are given in the table II below:

TABLE II
MEDICAL NAMED ENTITIES RULES EXAMPLE

CATEGORY	MEDICAL ENTITY EXAMPLE	RULE EXAMPLES
Symptom	Anorexie	(symptom name)
Clinical Review	Scanner AP	(test name)
Treatment	réhydratation 1 fl	(treatment name + ws + nbr + ws +unit)
Medication	Cefacidal 1g	(name drug+ws+nbr+unit)

V. EXPERIMENTAL STUDY

In this section we describe the data and metrics used to test our approach experimentally and discuss the different results.

A. Data set: clinical reports

We analyzed more than 50 French clinical reports to construct rules for medical named entities, and evaluated the system by using 30 new clinical reports from urology patients and general medicine at the hospital of CHLEF (Algeria). We have annotated the dataset with the help of a doctor. Five classes of medical entities were studied: Disease, Symptom, Treatment, Clinical review, Drug or medication. (so, 80 clinical reports have been collected in total: 50 for the development of rules and 30 for the evaluation of the system)

B. Metrics

Standard metrics for evaluating named-entity extraction are used to measure the accuracy of the proposed approach. We calculate precision, recall, and F-measure. They are defined as:

- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)
- F - Measure = 2 * (Precision * Recall)/(Precision + Recall)

Where:

- TP: True Positives; number of medical entities that were identified correctly.
- FP: False Positives; number of medical entities that were detected by the system and were not present in the report.
- FN: False Negatives; number of medical entities that were present in the report but system failed to detect them.

Table III describes in more details those metrics.

TABLE III
EVALUATION METRICS

		EXPERT (DOCTOR)	
		YES	NO
SYSTEM	YES	TP	FP
	NO	FN	TN

C. Experimental Results

In this study, we experiment the approach we have described in section 3 to recognize medical entities from clinical reports. Five categories were studied and the results are discussed in this section.

Fig. 6 shows the precision, recall and F-measure for each class. Analysis of the experiments allowed us to observe that the overall performance of our system over the five categories is good. The results are shown in table IV below.

The insufficient coverage of the diversity of all medical entities in our small set of rules explains the low results in recall. The system failed to recognize entities due to the insufficient numbers of entries in dictionnaries and insufficient

rules for identifying different entities especially, treatment and drugs entities.

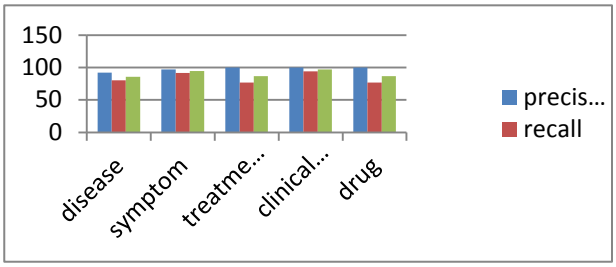


Fig.6. Performance system

Generally, the system performs well achieving and it gives a macro precision of 97,84% and a macro-recall of 83,78% which are the averge as it's shown in table V.

TABLE V
MACRO-AVERAGE RESULTS FOR PRECISION (P), RECALL (R) AND F-MEASURE (F) OF OUR SYSTEM.

EVALUATION	P %	R %	F %
AVERAGE	97,84	83,78	90,06

These results are very interesting and need to be evaluated in a larger collection of clinical reports, and this is very important.

VI. CONCLUSION

The work done in this paper is an attempt to broaden the coverage for medical entity extraction by incorporating the French clinical reports.

We used a rule based approach relying to the local grammar to extract medical entities from French clinical reports. The experimentations show that the rule based approach allows obtaining a good precision, but having a disadvantage to require a great human efforts and a considerable time compared to the high variability and the complex structure of the clinical reports.

One of the most important obstacles in identifying medical entities is the high terminological variation in the medical domain. In other hand the evolution of entity naming such as new abbreviations, names for new diseases or drugs constitute obstacles which can limit the scalability of the local grammar approach. Also the main limitation of the approach is their lack portability which limits their extension to other medical domains.

We plan to extract medical entities by machine learning, starting from a collection of training examples; the idea is to study the features of positive and negative examples of medical entities to be extracted over a collection of annotated documents with the need of doctor and design rules that capture instances of a given type. Therefore the hybridization will be a performance evaluation for future work.

REFERENCES

[1] A. Ben abacha, P. Zweigenbaum, "A Hybrid Approach for the Extraction of Semantic Relations from MEDLINE Abstracts", In Computational Linguistics and Intelligent Text Processing, 12th International Conference, volume 6608 of Lecture Notes in Computer Science, pages 139-150, February 20-26, Tokyo, Japan, 2011.

[2] F.Barigou, B.Beldjilali, B. Atmani. Using a cellular automaton to extract medical information from clinical Reports. Journal of information processing system, 8(1), 2012, 67–84.

[3] H. N. Traboulsi, "Named Entity Recognition: A Local Grammar-based Approach", Ph.D. dissertation, Dept of Computing, Surrey Univ. Guild ford, U.K, 2006.

[4] T. Poibeau, "Boosting the robustness of a named entity recognizer", International Journal of Semantic Computing, 2009, 32(1), pp 77-98.

[5] D. Nadeau, S. Sekine, "A survey of named entity recognition and classification", journal of linguistic investigations, 2007, 30(1), p .3-26.

[6] M. Mohammed Oudah, K. Shaalan, "A pipeline Arabic Named Entity Recognition Using a Hybrid Approach", in proceedings of COLING 2012, Mumbai: Technical Papers, pp 2159–2176.

[7] S. Meystre, G. Savova, K. Kipper-Schuler, J. Hurdle, "Extracting Information from Textual Documents in the Electronic Health Record: A Review of recent Research", year book of Medical Informatics. 2008, pp. 44-128.

[8] Y. He, M. Kayaalp. "Biological entity recognition with Conditional Random Fields.", In AMIA Annu Symp Proc, pp 293-297, 2008.

[9] F. Barigou, B. Beldjilali, B. Atmani, "MedIX : A Named Entity Extraction Tool from patient clinical reports", International Conference on Communication, Computing and Control Application, Hammamet, Tunisia, March 3-5, 2011, pp.488-494 .

[10] M. Chau, J., Xu, H. Chen, "Extracting Meaningful Entity from Polices Narrative Reports", Proceeding of the National Conference for Digital Government Research, 2002, pp.271-275

[11] L. Kosseim, G. Lapalme, "EXIBUM: un système expérimental d'extraction d'information bilingue", Rencontre International sur l'extraction, le filtrage et le résumé automatique (RIFRA'98), 1998.

[12] K. Shaalan, "Person Name Entity Recognition for Arabic", Proceedings of the 5th workshop on important Unresolved Matters, p 24-17, 2007.

[13] C. Friedman, P. Alderson, J. Austin, J. Cimino, S. Johnson, "A general natural language text processor for clinical radiology", Journal of the American Medical Informatics Association, 1994, 1(2), pp.161-174.

[14] P. Haug, L. Christensen, M. Gundersen, B. Clemons, S. Koehler, K. Bauer, "A natural language parsing system for encoding admitting diagnose ", American Medical Informatics Association Annual Symposium, AMIA 97, 1997, pp.814-818.

[15] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Meta thesaurus: the MetaMap program", American Medical Informatics Association Annual Symposium, AMIA'01, Washington, DC, USA, 2001, pp.17-21.

[16] A. Ben Abacha, P. Zweigenbaum, "Medical entity recognition: A comparison of Semantic and Statistical Methods", In Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT, pages 56–64, Portland, Oregon, USA, June 23-24.

[17] I. Spasic, F. Sarafraz, J. Akeane, G. Nenadic, "Medication information extraction with linguistic pattern matching and semantic rules", Published by group.bmj.com, 2010.

[18] M. Embarek, O. Ferret, "Learning patterns for building resources about semantic relations in the medical domain", Proceedings of the International Conference on Language Resources and Evaluation, LREC'08, Marrakech, Morocco, 26 May - 1 June, 2008.

[19] H. Harkema, R. Ian, R. Gaizauskas, M. Hepple (2005). Information Extraction from Clinical Records. In Proceedings of the 4th UK e-Science All Hands Meeting http://www.allhands.org.uk/2005/proceedings/,2005.

[20] C. A. Knirsch, N. Jain, A. Pablos-Mendez, C. Friedman, G. Hripcsak, "Respiratory Isolation of Tuberculosis Patients Using Clinical Guidelines and an Automated Clinical Decision Support System". Journal Infection Control and Hospital Epidemiology, 1999, 19(2), pp.94-100.

[21] T. Sibanda, T. He, P. Szolovits, O. Uzuner, "Syntactically-informed semantic category recognition in discharge summaries", Proceeding of the Fall Symposium of the American Medical Informatics Association; Washington, DC, November, 2006.

[22] S. Zhang, N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts", Journal of Biomedical Informatics 46, 2013, p 1088-1098.

[23] J. Fan, N. Sood, Y. Huang. "Disorder Concept Identification from Clinical Notes An Experience with the ShARe/CLEF 2013 Challenge", Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, 2003, Valencia - Spain.

[24] S. Matos, T. Nunes, J. L. Oliveira. "BioinformaticsUA: Concept Recognition in Clinical Narratives Using a Modular and Highly Efficient Text Processing Framework", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, August 23-24, 2014, pages 135–139.

[25] S. Ramanan, S. Nathan. "RelAgent: Entity Detection and Normalization for Diseases in Clinical Records: a Linguistically Driven Approach", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 477–481, Dublin, Ireland, August 23-24, 2014.

[26] Y. Xia, X. Zhong, P. Liu, C. Tan, S. Na, Q. Hu and Y.Huang. " Combining MetaMap and cTAKES in Disorder Recognition: THCIB at CLEF eHealth Lab 2013 Task 1", OnlineWorking Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 - 26 September, 2013, Valencia -Spain.

[27] J. D. Osborne, B. Gyawali, T. Solorio. "Evaluation of YTEX and MetaMap for clinical concept recognition", Online Working Notes of the CLEF 2013 Evaluation Labs andWorkshop, 23 - 26 September, 2013, Valencia - Spain.

[28] P. Pathak, P.Patel, V.Panchal, N. Choudhary, A. Patel, G. Joshi. "ezDI: A Hybrid CRF and SVM based Model for Detecting and Encoding Disorder Mentions in Clinical Notes", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 278–283, Dublin, Ireland, August 23-24, 2014.

[29] K. Gojenola, M.Oronoz, A. Pérez, A. Casillas. " IxaMed: Applying Freeling and a Perceptron Sequential Tagger at the Shared Task on Analyzing Clinical Texts", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 361–365, Dublin, Ireland, August 23-24, 2014.

[30] A. Bodnari, L. Deleger, T. Lavergne, A. Neveol, P. Zweigenbaum. "A Supervised Named-Entity Extraction System for Medical Text", Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 – 26 September, Valencia - Spain.

[31] A. Parikh ,Ah PVS, J. Mustafi, L. Agarwalla, A. Mungi. "ThinkMiners: Disorder Recognition using Conditional Random Fields and Distributional Semantics", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 652–656, Dublin, Ireland, August 23-24, 2014.

[32] Y. Zhang, J.Wang, B.Tang, Y.Wu, M. Jiang, Y. Chen, H. Xu. "UTH_CCB: A Report for SemEval 2014 – Task 7 Analysis of Clinical Text", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 802–806, Dublin, Ireland, August 23-24, 2014.

[33] G.Attardi, V. Cozza, D.Sartiano. "UniPi: Recognition of Mentions of Disorders in Clinical Text", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 754–760, Dublin, Ireland, August 23-24, 2014.

[34] J.Jonnagaddala, M. Kumar, H.J. Dai, E. Rachmani, C.Y. Hsu. "TMUNSW: Disorder Concept Recognition and Normalization in Clinical Notes for SemEval-2014 Task 7", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 663–667,Dublin, Ireland, August 23-24, 2014.

[35] G. Omid, R.J. Kate. "UWM: Disorder Mention Extraction from Clinical Text Using CRFs and Normalization Using Learned Edit Distance Patterns", Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 828–832, Dublin, Ireland, August 23-24, 2014.

[36] J.Cogley, N. Stokes, J. Carthy. " Medical Disorder Recognition with Structural Support Vector Machines", Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, 23 – 26 September, Valencia - Spain.

[37] C. Grouin, P. Zweigenbaum, "Automatic de-identification of French clinical record: comparison of rule based and machine learning approaches", In Proc MEDINFO 2013, Studies in Health Technology and Informatics, pages 476-480. Amsterdam, IOS Press, 2013.

[38] S. Sarawagi, "Information extraction. Foundations and Trends in Databases". (2007). Vol. 1, No. 3. 261–377.

[39] J. Jiang," Information Extraction from Text". Research Collection School of Information Systems. In Charu C. Aggarwal and ChengXiang Zhai (Eds.), (2012). Mining Text Data, Springer. 11-41.

[40] H. Ware, J. M. Charles, J. Vasudevan, R. Oussama. "Machine learning-based coreference resolution of concepts in clinical documents". (2012). J Am Med Inform Assoc; 19:883e887. doi:10.1136/amiajnl-2011-000774.

[41] W.Sun, A. Rumshisky , & O. Uzuner, "Evaluating temporal relations in clinical text: 2012 i2b2 Challenge". (2013). In the Journal of the American Medical Informatics Association. doi:10.1136/amiajnl-2013-001628.

[42] J. R. Hobbs, D.Appelt, M. Tyson, J. Bear, and D. Islael, "SRI International: Description of the FASTUS system used for MUC-4".(1992). In Proceedings fo the 4th Message Understanding Conference (MUC-4), 268–275.

[43] G. Krupka,P. Jacobs, L.Rau, L. Childs, and I.Sider,"GE NLTOOLSET: Description of the system as used for MUC-4". (1992). In Proceedings of the 4th Message Understanding Conference (MUC-4), 177–185.

[44] D. Ayuso, S.Boisen, H. Fox, H .Gish, R .Ingria, and R .Weischedel,.(1992). "BBN: Description of the PLUM system as used for MUC-4". In Proceedings of the Fourth Message Understanding Conference (MUC-4), 169–176.

[45] Yangarber, R. & Grishman, R.(1998). NYU: Description of the Proteus/PET system as used for MUC-7 ST. In Proceedings of the 7th Message Understanding Conference: MUC-7, Washington, DC.

[46] Kaiser, K., & Miksch, S.(2005). "Information Extraction.A Survey.Vienna University of Technology".Asgaard-TR-2005-6.

[47] H Cordobés,., A. Fernández Anta, L. F. Chiroque, F. Pérez, T. Redondo, and A. Santos, "Graph-based Techniques for Topic Classification of Tweets in Spanish", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 2, issue Special Issue on AI Techniques to Evaluate Economics and Happines, no. 5, pp. 31-37, 03/2014.

[48] K. Khan,and A. Sahai, "A fuzzy c-means bi-sonar-based Metaheuristic Optimization Algorithm", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 1, issue Regular Issue, no. 7, pp. 26-32, 12/2012.

[49] C. Chang, M. Kayed, M.R. Girgis, K. Shalan, "A survey of web Information Extraction Systems".(2006). IEEE transactions on knowledge and data engineering, TKDE-0475-1104.R3.

[50] H .Gurulingappa, A. Matteen-rajput, & L. Toldo, "Extraction of Adverse Drug Effects from Medical case Rets". (2012). In: Courtot M, editor. International Conference Biomedical Ontologies, 22-25. Graz, Austria.

[51] H.Bolivar-Baron,., R. Gonzalez-Crespo, and O. Sanjuan-Martinez, "Ontology of a scene based on Java 3D architecture.", International Journal of Interactive Multimedia and Artificial Inteligence, vol. 1, issue Special Issue on Business Intelligence and Semantic Web, no. 2, pp. 14-19, 12/2009.

[52] Z. Harris, "Theory of language and Information: A Mathematical Approach", Oxford & New York: Clarendon Press, 1991

[53] H. N. Traboulsi, "Arabic Named Entity Extraction: A Local Grammar–based Approach", Proceeding of the International Multiconference on Computer Science and Information Technology, 2009, pp. 139-143.

[54] M. Gross, "The construction of local grammars", in E.Roche & Y. Schabés (eds), Finite-State Language, Speech, and communication, MIT Press, 1997, pp.329-354.

[55] S. J. Bolaños-Castro, R. G. Crespo, V. H. Medina-García, "Patterns of software development process", International Journal of Interactive Multimedia and Artificial Intelligence, vol 1. Issue 4, pp. 33-40, 12/2011

Aicha Ghoulam graduated from Department of Computer Science, University of Chlef, Algeria. In 2010, she received his Magister degrees in Computer Science from Algiers University. She is currently a research member of Laboratory of Computer Science of Chlef. Her research interests include natural language processing, information extraction, information retrieval, knowledge-based system, pattern recognition.

Fatiha Barigou graduated from Department of Computer Science, University of Oran, Algeria. In 2012, she received his PhD degrees in Computer Science from the University of Oran. Dr. Barigou is currently a research member of Laboratory of Computer Science of Oran. Her research interests include natural language processing,

information extraction, information retrieval, knowledge-based system, pattern recognition and data mining.

Ghalem Belalem. Graduated from Department of computer science, Faculty of exact and applied sciences, University of Oran, Algeria, where he received PhD degree in computer science in 2007. His current research interests are distributed system; grid computing, cloud computing and data grid placement of replicas, consistency, fault tolerance, economic models, energy consumption, Big data, and improved performance in large scale systems and mobile environment.

Farid Meziane Professor in Data and Knowledge Engineering. He holds a PhD in Computer Science from the University of Salford and is the head of the Informatics Research Centre. He is the Chair of the 20th International Conference on Application of Natural Language to Information Systems (NLDB2015) and has served on the programme committees of over 20 conferences. He is on the editorial board of 5 international journals. His research interests are in the broad area of data and knowledge engineering. This includes data mining, information extraction and retrieval, big data and the semantic web.

OpinAIS: An Artificial Immune System-based Framework for Opinion Mining

Alejandro Baldominos Gómez, Nerea Luis Mingueza, Mª Cristina García del Pozo

Universidad Carlos III de Madrid, Leganés, Spain

***Abstract* — This paper proposes the design of an evolutionary algorithm for building classifiers specifically aimed towards performing classification and sentiment analysis over texts. Moreover, it has properties taken from Artificial Immune Systems, as it tries to resemble biological systems since they are able to discriminate harmful from innocuous bodies (in this case, the analogy could be established with negative and positive texts respectively). A framework, namely OpinAIS, is developed around the evolutionary algorithm, which makes it possible to distribute it as an open-source tool, which enables the scientific community both to extend it and improve it. The framework is evaluated with two different public datasets, the first involving voting records for the US Congress and the second consisting in a Twitter corpus with *tweets* about different technology brands, which can be polarized either towards positive or negative feelings; comparing the results with alternative machine learning techniques and concluding with encouraging results. Additionally, as the framework is publicly available for download, researchers can replicate the experiments from this paper or propose new ones.**

***Keywords* — Artificial immune system, evolutionary computation, sentiment analysis, machine learning, classification.**

I. Introduction

SENTIMENT ANALYSIS (also referred as *opinion mining*) [27] is a field of Natural Language Processing (NLP) which aims at extracting emotional or subjective information from a source, which may be a document, a website, a publication in a social network, etc.

A specific task within sentiment analysis is retrieving the polarity of the document, i.e., whether it expresses a positive or negative feeling (sometimes, the case when the document does not express any feeling at all is also observed). This is definitely not a simple task, as natural language semantics are very complex, and there are many ways, sometimes too rhetorical, to express a positive or negative feeling. In fact, sentiment analysis involves so many challenges that many works over the last decade have discussed them [28, 22, 24, 38] and most if not all of those difficulties remain invariant and are widely discussed today [3], as social networks start to set up enormous corpus which are increasingly interesting for this task [26, 19].

From the computational side, a Machine Learning (ML) approach perfectly fits this task. The problem of guessing the polarity of a document is analogous to a binary classification problem. Yet some decisions, such as how the features for classification are retrieved from the document, or which particular ML algorithm will be used must be taken before some results could be obtained.

This work aims at applying an Artificial Immune System (AIS) approach, which is a biologically-inspired ML technique based on the immune system of vertebrates, to solve this problem. Actually, the algorithm can be easily extended to support multiclass classification and prediction problems. To provide additional value, this work also have the purpose of developing a framework which can be extended to new algorithms and applications, so that it can be reused by the scientific community.

A brief introduction of AIS, as well as some related work is provided in section 2. Sections 3 and 4 discuss how features can be extracted from text, and how the immune-based algorithm is applied for the sentiment analysis task. Meanwhile, section 5 focuses on the design and the development of OpinAIS, the AIS-based framework for solving sentiment analysis problems.

Finally, section 6 shows some results obtained from using OpinAIS with two public datasets involving voting records for the US Congress and a Twitter corpus. Secion 7 provides some conclusive remarks on this work, and appendices are included which detail how to run and extend the framework.

II. Related Work

The purpose of this section is to briefly discuss the problem of sentiment analysis, analysing previous works where ML techniques were used to face this problem, and finally describing how AIS work and some state of the art applications where AIS are applied for the task of sentiment analysis.

As it was stated in the previous section, sentiment analysis (or opinion mining) is a problem that involves many of the challenges brought by NLP. Techniques located within the field of Artificial Intelligence (AI) are well suited for facing this problem [5]. In particular, detecting the polarity of a text (whether it contains positive or negative feelings) can be in most cases reduced to a problem of binary classification by using bag of words (where binary attributes indicate whether a particular word appear or not in the document). By doing so, many classical ML techniques can be applied [18], including

DOI: 10.9781/ijimai.2015.333

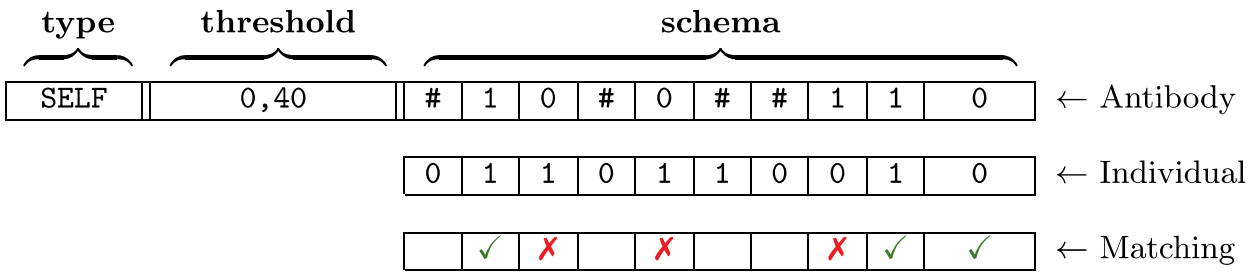


Fig. 1. Example of an antibody detecting an individual.

variations of the Naïve Bayes classifier [40], Support Vector Machines [37], Kernel Trees [1] or semi-supervised approaches [7, 21]. Some works compare several of these techniques for detecting emotions and personality in

platforms such as Whatsapp or SMS [33]. Other works do not use bag of words but rather different approaches such as graph-based techniques for *tweet* classification [6], while in this case the work do not focus on opinion mining but rather topic detection. In the recent years, several surveys summarizing the most relevant techniques and contributions have been published [23, 25, 42, 16].

Biologically inspired AI techniques have also proved to be relevant for solving this problem, as many works in the last couple of years use these approaches for opinion mining, such as it is the case of genetic algorithms [14], particle swarm and ant colony optimization [39], neural networks [2] or a combination of several of them [17].

Regarding biologically-inspired artificial intelligence, relevant techniques include AIS which appeared in the mid-90s, when efforts for understanding the immune system [12] significantly increased. The idea beyond these systems is to imitate the biological adaptive immune system and its ability to recognize external harmful individuals, which can be generalized to approach and solve a variety of problems.

The present work is based on a previous research on applying AIS to document classification [41], which was already based on an earlier work that applied AIS-based techniques for concept learning [30]. In these approaches, which will be described in further detail in section 4, a population of antibodies is evolved with a co-evolutionary technique. Eventually, a set of antibodies conform a classifier, which can be used to infer the class a certain item. As long as a document classification problem can be represented as a binary string, the system can learn a classifier from a set of training instances. Additionally, there are more recent works which study the convenience of using AIS for opinion mining [34], and use this kind of techniques for selecting features for opinion mining [35] or analysing sentiments in newspapers [31].

The use of an evolutionary algorithm somehow recalls from other AIS techniques such as clonal selection, as the evolutionary operators resemble the operators in algorithms such as *CLONALG* [8]. Additionally, the process of affinity maturation is achieved by the evolutionary algorithm, which tries to increase the fitness of the antibodies, i.e., their ability to correctly detect antigens.

Finally, prior work proposed a theoretical framework for AIS [8], and other ML frameworks such as Weka [15] also incorporates AIS-based techniques for general-purpose classification as well as specific text mining algorithms which would enable performing document classification or opinion mining [32]. However, the framework proposed in this work is more specifically aimed towards document classification (and opinion mining in particular) and therefore is simpler to be used and to be extended, whereas others are more complete and supports other problems beyond classification itself but fail to provide such specific parameterization for opinion mining.

III. Data Wrangling

For applying ML to documents expressed in natural language, a preliminary phase of data wrangling is often required so that these can be converted to a format accepted by the algorithm. For this work, the input is converted into a binary string (a list of boolean features).

The current section details the process followed to obtain a set of binary individuals from a set of documents expressed in natural language.

A. Preprocessing

When dealing with natural language, some processing of the input may lead to better results, as raw data is typically too noisy. An approach to this processing involves implementing a series of filters running in a pipeline [11], each of those performing some processing over data, which is then inputted to the next filter. The ultimate goal of these phases is to increase the ratio of meaningful words by reducing the total number of different words, while trying to keep semantics. This section describes the preprocessing phase applied in this work, and how it could help to improve the results.

1) **Removing Non-Alphanumerical Symbols**: usually, non-alphanumerical symbols in a text lack from any semantic meaning thus can be ignored. However, other words formed only by symbols (e.g. emoticons such as :-) or xD) not only do have semantic meaning, but also store a strong emotional load [20].
2) **Converting to Lower Case**: in many cases, words keep semantics regardless whether they are written uppercase or lowercase. For this reason, it is useful to turn all symbols to the same case, to represent the same word always with the same characters.

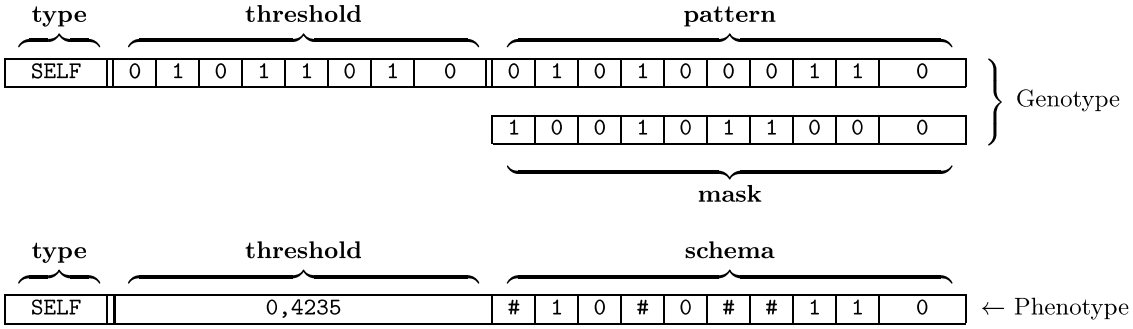


Fig. 2. Example of a translation between an antibody's genotype and phenotype

3) **Removing Stop Words**: in natural languages, there are many words that are completely meaningless, and only have syntactical value, such as determiners, prepositions, etc. These words usually appear with very high frequencies, and so may lead the algorithm to think that they are relevant. While there are works involving the automatic identification of stop words [44], the approach followed in this work uses a stop words dictionary for the English language.
4) **Stemming Words**: stemming is the process to reduce a word to its stem (e.g. "work", "working" and "worked" are all reduced to "work"). This way, the space of words is considerably reduced, while the original meanings persist, increasing the ratio of semantics versus the number of different tokens. This work uses the Porter Stemming Algorithm [29], as it is widely used and considered the *de facto* standard for stemming English words.

B. Extracting the Features

After the preprocessing phase, the input is still a set of documents, each of these reduced to a set of tokens (stems) resulting from applying the filters to the original words. The purpose of the second step is to decide which of the tokens are more relevant when deciding the class of each document. To do so, a metric known as *expected information gain* may be used, which estimates the information provided by a token based on the entropy of the set of documents containing and not containing that token.

In particular, the expected information gain for a word w and a set of documents S is calculated as follows:

$$E(w,S) = I(S) - \left(P(w)I(D_w)\right) + \left(P(\neg w)I(D_{\neg w})\right)$$

where:

- $P(w)$ is the probability that w appears in a document, i.e., the percentage of documents containing w.
- $P(\neg w)$ is the probability that w does not appear in a document, i.e., the percentage of documents not containing w.
- $D(w)$ is the subset of documents containing w.
- $D(\neg w)$is the subset of documents not containing w.
- $I(S), S = \{D, D_w, D_{\neg w}\}$ is the entropy of the set S for each of the classes, which is defined as follows:

$$\mathrm{I(S)} = \sum_{\mathrm{c}\in\{+,-\}} -\mathrm{P(S_c)}\log_2(\mathrm{P(S_c)})$$

The computation of the entropy can only be performed if a training set exists where the class is known in advance for each document in the set, i.e., under a supervised learning scheme. While the previous equation refers to the class as either positive or negative, it could be generalized to any arbitrary number of different classes.

Finally, when the expected information gain is computed for all words, the n words with the highest value of E are chosen, which can be expressed as $F = \{w_1\ w_2\ w_3\ ...\ w_n\}$.

During the last phase, known as vectorization, the objective is to convert documents to individuals represented by a binary string: $d = \{b_1\ b_2\ b_3\ ...\ b_n\}$. To do so, for each bit b_i in the individual representing the document, $b_i = 1$ if the document contains the word w_i, or $b_i = 0$ if it does not. After the vectorization process takes place, the original set of documents is converted into a set of binary individuals.

IV. The Artificial Immune System

For the development of AIS, the approach provided in [41] is followed. This section provides first an intuition of how antibodies are represented in the AIS and how they can be used to detect individuals of a certain type. Later, it describes the process to obtain a classifier, composed of a set of antibodies, given a set of training examples.

A. Design of Antibodies

The design of antibodies is a key task in the development of an AIS, as they are the entities responsible for detecting the type of the individuals and, in the end, of the classification task. In the AIS developed for this work, antibodies (also called *detectors*), integrate the next elements:

- A **type** indicating the polarity of the individuals this detector should recognize, which usually are *self* (i.e., part of the body) or *non-self* (i.e., foreign to the body and thus potentially harmful), which in the task of sentiment analysis are identified to positive and negative items respectively. Nevertheless, this definition can be extended to support a set of k different classes.

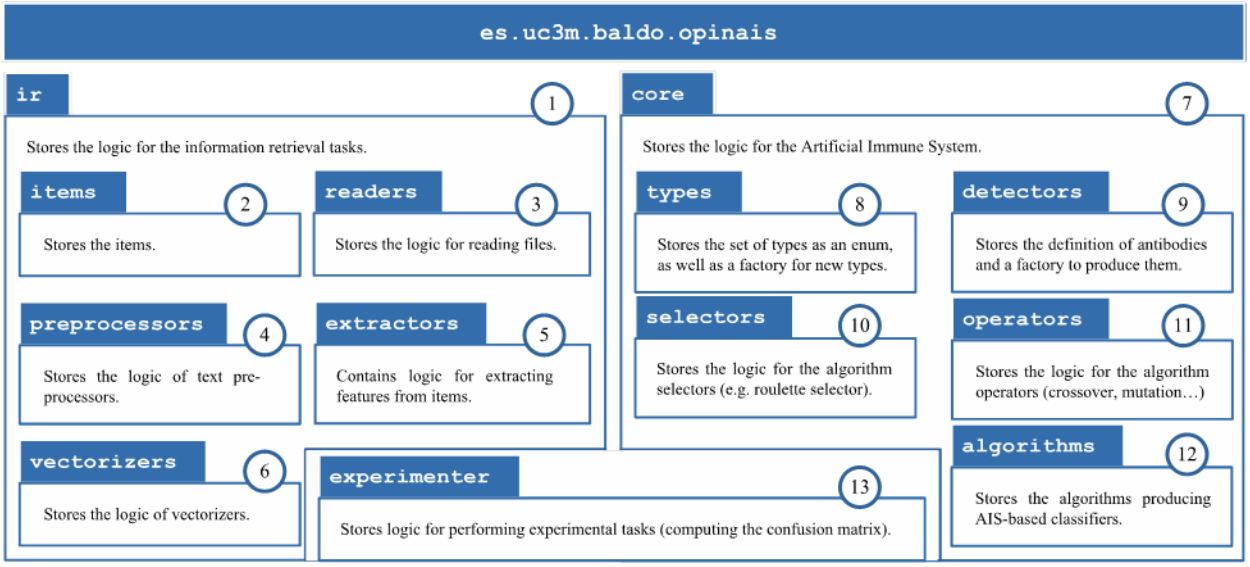


Fig. 3. Package structure for the OpinAIS framework

- A binary **schema** of the same length that the number of features of an individual. While this schema is binary, besides *0*s and *1*s it can also contain wildcard (*#*) positions.
- A real threshold in the interval [0,1].

After a representation for the antibodies is chosen, it is important to decide the process by which an antibody detects an individual as being from its type. The steps for this process are detailed next:

1) Each bit in the schema is compared to each bit of the individual in the same position. Bits in wildcards positions, marked as *#*, are ignored.
2) The matching ratio is computed as the number of bits matching in the same position divided by the total number of comparisons performed (i.e., the number of non-wildcard positions in the antibody's schema).
3) If the matching ratio exceeds the threshold, then the antibody detects the individual as being of its same type. Otherwise, it does not detect it.

An example of individuals matching is provided in Figure 1. In this particular example, there are 3 matches from a total of 6 comparisons, so the matching ratio is $3/6 = 0.5$, which is greater than the threshold ($0.5 > 0.4$), and thus the individual is recognized as *self*.

B. Evolving the Classifier

A classifier is a set of antibodies, one for each possible type. When an individual is inputted to the classifier, each antibody tries to detect it. The type of the individual is obtained as the type of the antibody who detects the individual and maximizes the matching ratio. It is important to notice that, in the case that no antibody detects the individual, then it remains unclassified.

An evolutionary approach is chosen to obtain the classifier. Before the details of the algorithm are described, it is important to decide the way the antibody is represented in order to be treated by the evolutionary algorithm, i.e., its genotype. Antibodies are encoded as follows:

- The **type** does not need to be encoded, as the evolutionary operators do not affect it.
- The **threshold** is encoded as an 8-bit number in Gray code, as small changes in this binary representation lead to small changes in the number it represents. Because an 8-bit string represents an unsigned integer n the interval [0, 255], the resulting value is normalized in the range [0, 1], thus dividing it by 255.
- The **schema** is represented by two different binary strings, named **pattern** and **mask**, both with the same length that the schema. Given a pattern and a mask, the schema can be determined as follows:
  1. If the *i*-bit in the mask is *1*, then the corresponding bit in the schema will be a wildcard (*#*).
  2. If the *i*-bit in the mask is *0*, then the corresponding bit in the schema will correspond to the *i*-bit from the pattern.

  It must be noticed that with this encoding, many different genotypes may translate into the same phenotype. Actually, this is common in natural immune systems, as different chains of amino acids may fold into antibodies recognizing the same pattern [30].

An example of a translation between the genotype and the phenotype is shown in Figure 2. Once the binary representation for the antibodies is depicted, the details of the evolutionary algorithm can be discussed. This algorithm follows the next steps:

1) **Initialization**: to begin with, the algorithm generates an initial population of antibodies, of a fixed (yet configurable) size. While this initialization is performed randomly, it attends to some parameters:
   - The **type bias** represents the probability that the new antibody detects *self* individuals. For instance, if the type bias has a value of 0.6, then it means that in average, 60% of the antibodies in the population will detect *self* individuals.

- The **generality bias** represents the probability that a bit in the schema is a wildcard (#). For instance, if the generality bias has a value of 0.3, then it means that in average, 30% of the bits in an antibody schema will be wildcards.

2) **Fitness calculation**: once the initial population is generated, the algorithm calculates the fitness of each antibody. This fitness is calculated as the number of correctly classified individual minus the number of false positives. Unclassified individuals are considered as correctly classified if they are not of the same type that the antibody. To prevent negative values for the fitness, it is normalized in the range [0, 1].
3) **Selection**: two antibodies from the same type are chosen, in a random yet fitness-proportional manner. To do so, a technique known as *roulette selection* is performed, by which antibodies with higher fitness have more chances to be selected.
4) **Reproduction**: the two selected antibodies serve as parents for a new one. This reproduction is performed using *standard crossover*, by which the genome of the child antibody is filled by selecting, for each bit, one random bit in the same position from either of their parents. A parameter, known as the **crossover rate**, establishes the probability that crossover takes place. In the cases where crossover is not performed, the child results as an exact copy of one of their parents.
5) **Mutation**: the child is mutated, by performing bit flipping for each individual bit. In this case, a parameter known as **mutation rate** controls the probability that a single bit is flipped.
6) **Generational replacement**: steps 3-5 are performed until the new population has as many antibodies of the same type as the previous one. When such a thing occurs, the original population is replaced with the new one. During this phase, *elitism* can be introduced through a parameter, known as **elitism rate**, which controls the percentage of best antibodies that are kept between generations. By introducing elitism, the best detectors are maintained.
7) **Stop condition**: if the maximum number of generations is not achieved, the algorithm restarts from step 2. Otherwise, the algorithm stops and a classifier is built by choosing the best antibody from each type. Each bit in the schema is compared to each bit of the individual in the same position. Bits in wildcards positions, marked as #, are ignored.

An improved version of this algorithm, which includes a cooperative approach, has been also developed. This algorithm, which is known as *co-evolutionary algorithm*, inserts a new phase after the fitness calculation. Indeed, it performs a second fitness calculation, which computes a *cooperative fitness* rather than an individual one.

Particularly, the cooperative fitness for a certain antibody is the result of classifying all the individuals with that antibody combined with the best antibodies of the remaining types. The fitness sums up all the hits and subtracts the misses (i.e., individuals wrongly classified). It remains as a user-configurable parameter to decide whether unclassified individuals are considered as wrongly classified or are ignored (some applications may benefit from ignoring unclassified instances, such as those where unclassified instances are preferred over misclassified ones). Finally, the fitness is normalized in the range [0, 1].

The cooperative fitness evaluates a potential classifier rather than each antibody itself. For this reason, results are usually better, but computing time can also be significantly higher.

V. The OpinAIS Framework

OpinAIS is an extensible framework that enables the application of AIS to a variety of classification problems, as long as instances can be represented as binary strings.

The power of OpinAIS remains in its extensibility. While the algorithms described in the prior section are already implemented, it is relatively simple to develop new ones. This principle not only applies to algorithms, but also to input readers, information retrieval processors, etc.

Figure 3 shows the package structure of the framework. The heading of the figure refers to the path where the OpinAIS framework is placed in the package. The purpose of this section is to describe the responsibility of each one, so that it can serve as a quick developer guide. The *ir* package (1) stores the logic required to retrieve a set of individuals (which are computable by the algorithm) from an input source, such as a text file containing one individual in each line. This package is divided in several subpackages, with clearly defined responsibilities:

- Package *items* (2) contains items, which a generic type of individuals, i.e., *something* which can be potentially converted to an individual, but so far is not, such as a *tweet* or an HTML document are some kinds of items.
- Package *readers* (3) will store readers, whose responsibility is to generate a set of items from an input source. For instance, there may be a folder containing HTML documents, and a reader that returns a set of objects representing those.
- Package *preprocessors* (4) stores classes responsible for performing some preprocessing tasks over text items which may significantly increase the performance of the classifier, as it was shown in section 3.
- Package *extractors* (5) contains classes whose purpose is to extract features from a set of items in those cases when specific logic for this task is required (e.g. implementation details on the extractor for text items was provided in section 3).
- Package *vectorizers* (6) contains the logic for converting items into individuals encoded as a set of bits. The behaviour of vectorizers for text items was explained in section 3.

On the other hand, the *core* package (7) stores the logic required for obtaining a classifier from a set of input individuals. Most details on this process were already given in section 4, so this section will limit to explain how this

functionality is broken into different packages:

- Package *types* (8) contains an enumerated type, *Type* with the set of all possible classes for individuals. For the sake of flexibility, this enumerated type is empty and it is filled dynamically when classes are known.
- Package *detectors* (9) contains the *Detector* class, representing the definition of an antibody, as it was described in section 4. Moreover, the *DetectorFactory* class implements some logic for initializing the first population of antibodies.
- Package *selectors* (10) contains classes which implement some logic for choosing an antibody from a population. So far, the evolutionary algorithm developed uses a roulette selector, yet many others could be implemented by the user (e.g. a tournament selection).
- Package *operators* (11) stores auxiliar logic containing operators used by the algorithms. An example of such operators is the standard crossover and mutation, which were already described in section 4, and which are implemented in the classes *CrossoverOperator* and *MutationOperator* respectively.
- Package *algorithms* (12) stores the algorithms, whose responsibility is to receive a set of training individuals and build a classifier from them.

Finally, the *experimenter* package (13) contains additional logic for assisting the experimental tasks, such as:

- dividing a set of individuals into training and test sets, where the size of these tests can be set by the user.
- given a classifier and a set of individuals, computing the confusion matrix, i.e., a table showing up the number of correctly classified instances, as well as false positives, true negatives and unclassified individuals.
- computing the performance of a classifier measured as its accuracy, given the corresponding confusion matrix.

The OpinAIS framework is publicly available for download from a GitHub repository[1], and more information and developers documentation can be found in the project website[2], including instructions on how to run and extend it.

VI. Experimental Results

Once the OpinAIS framework is developed, experiments over two different datasets are performed using the implemented AIS-based algorithm in order to validate the system, and a comparative evaluation with classic ML techniques is also carried out. Additionally, further evaluation for the algorithms underlying this proposal has already been published in previous works [30, 41].

A. US Congressional Voting Records Dataset

The first experiment in this section will execute over the *US Congressional Voting Records Data Set* from the *UCI Machine Learning Repository* [36], which is composed of 117 (34.21%) instances of republican votes and 225 (65.79%) instances of democrat votes. As instances are fairly unbalanced, two experiments are executed: the first one will deal with all input instances, while the second will balance them, thus taking 117 instances of republican congressmen and the same number of democrats.

TABLE I
Confusion matrix for the US Voting Records dataset with unbalanced data

	Republican	Democrat	N/C
Republican	86/22	5/3	1/0
Democrat	2/2	179/41	1/0

TABLE II
Confusion matrix for the US Voting Records dataset with balanced data

	Republican	Democrat	N/C
Republican	85/26	3/0	2/1
Democrat	2/0	95/19	1/0

The results for an average execution are shown in Table I and Table II, which contains the confusion matrix for the experiment with unbalanced and with balanced data respectively. Cells in the confusion matrix contain two values, the first one referring to the training set and the second to the test set. The last column refers to non-classified instances.

As it can be seen, the results are pretty good. For the original dataset, the accuracy is 96.72% (92.65%), while for the balanced subset the value is 95.74% (97.83% for the test set). From these numbers, two conclusions can be drawn: in the first place, the AIS-based evolutionary algorithm provides good classification accuracy, which validates that the system is working properly. Secondly, the algorithm shows a good generalization ability, as long as it does not fall into overfitting the individuals from the training set, achieving very similar results for both the training and the test sets.

B. Twitter Sentimental Corpus

The second battery of experiments is performed over a set of actual *tweets*, which are short publications in Twitter. This dataset is provided by Sanders Analytics[3], consists in 5,513 tweets about technological companies and it is especially interesting as it provides two different classifications. The first one has to do with the polarity of the *tweet*, which can either be *positive*, *negative*, *neutral* or *irrelevant*. In most cases, *tweets* are classified as *irrelevant* if they are written in languages other than English or have nothing to do with the topic (i.e., they are spam). Also, *tweets* are classified as *neutral* when they are neither positive nor negative in a clear way, they are simple factual statements or they express questions with no strong emotions. The second classification criterion has to do with the technological enterprise related to the content of the *tweet*.

An example of a *tweet* for each polarity and enterprise is

[1] http://github.com/alexbaldo/opinais
[2] http://baldo.uc3m.es/opinais/
[3] http://www.sananalytics.com/lab/twitter-sentiment

shown in Table III. As it can be seen, the classification task may find some difficulties. In the first place, a *tweet* with negative polarity contains a high load of positive words such as "greatly impressed". Secondly, as irrelevant *tweets* can be in any language, they add a huge number of possible words that may harden the features extraction task.

TABLE III
TWEETS FOR EACH POLARITY AND ENTERPRISE

Brand	Polarity	*Tweet*
Apple	Positive	@apple @siri is efffing amazing!
Microsoft	Neutral	Creating #Pareto charts using #Microsoft #Excel
Google	Negative	Not greatly impressed with #Google and #Samsung presentation skills.
Twitter	Irrelevant	#twitter sos un vicioooooo

TABLE IV
CONFUSION MATRIX FOR THE TWITTER SENTIMENT CORPUS WHEN CLASSIFYING THE *TWEET'S* ENTERPRISE USING OPINAIS

	A	G	M	T	NC
A	699/75	42/2	27/10	44/3	165/29
G	27/4	742/68	59/5	96/12	196/25
M	50/6	59/6	791/93	24/1	208/19
T	20/4	18/2	82/9	825/93	128/11

TABLE V
CONFUSION MATRIX FOR THE TWITTER SENTIMENT CORPUS WHEN CLASSIFYING THE *TWEET'S* POLARITY USING OPINAIS

	Positive	Negative	NC
Positive	296/30	107/20	42/1
Negative	72/9	385/39	32/4

TABLE VI
RESULTS FOR THE EVALUATION OF THE TWEET'S POLARITY PREDICTION USING OPINAIS

	Train	Test		Train	Test
ACC	72.9%	67.0%	TPR	73.5%	60.0%
PPV	80.4%	76.9%	TNR	84.3%	81.3%
NPV	78.3%	66.1%	MCC	58.2%	48.9%

TABLE VII
CONFUSION MATRIX FOR THE TWITTER SENTIMENT CORPUS WHEN CLASSIFYING THE *TWEET'S* POLARITY USING NAIVE BAYES

	Positive	Negative
Positive	196/80	144/76
Negative	47/16	339/139

For both experiments, a population of 200 individuals each one recognizing 250 features is evolved during 1000 generations. The dataset is divided into a training set containing 70% of the original instances and a test set with the remaining 30%.

First, a classifier is trained to infer the enterprise related with the *tweet*, either Apple (A), Google (G), Microsoft (M) or Twitter (T). The number of instances for each enterprise is approximately the same, i.e., the dataset is balanced. The confusion matrix for this problem is shown in Table IV. As it can be seen the results are good, providing a great improvement over random guess with an accuracy of 92.41% (90.41% over the test set, which also shows a good generalization).

TABLE VIII
CONFUSION MATRIX FOR THE TWITTER SENTIMENT CORPUS WHEN CLASSIFYING THE *TWEET'S* POLARITY USING C4.5

	Positive	Negative
Positive	233/90	107/66
Negative	50/37	336/18

TABLE IX
CONFUSION MATRIX FOR THE TWITTER SENTIMENT CORPUS WHEN CLASSIFYING THE *TWEET'S* POLARITY USING RANDOM FOREST

	Positive	Negative
Positive	317/126	23/30
Negative	29/34	357/121

TABLE X
CONFUSION MATRIX FOR THE TWITTER SENTIMENT CORPUS WHEN CLASSIFYING THE *TWEET'S* POLARITY USING SVM

	Positive	Negative
Positive	143/61	197/5
Negative	40/12	346/143

TABLE XI
CONFUSION MATRIX FOR THE TWITTER SENTIMENT CORPUS WHEN CLASSIFYING THE *TWEET'S* POLARITY USING MLP

	Positive	Negative
Positive	198/79	142/77
Negative	26/13	360/142

TABLE XII
CONFUSION MATRIX FOR THE TWITTER SENTIMENT CORPUS WHEN CLASSIFYING THE *TWEET'S* POLARITY USING CLONALG

	Positive	Negative
Positive	181/81	159/75
Negative	78/37	308/118

TABLE XIII
CONFUSION MATRIX FOR THE TWITTER SENTIMENT CORPUS WHEN CLASSIFYING THE *TWEET'S* POLARITY USING AIRS

	Positive	Negative
Positive	145/65	195/91
Negative	44/15	342/140

TABLE XIV
CONFUSION MATRIX FOR THE TWITTER SENTIMENT CORPUS WHEN CLASSIFYING THE *TWEET'S* POLARITY USING IMMUNOS-81

	Positive	Negative
Positive	113/55	227/101
Negative	36/11	350/144

Finally, the algorithm is executed with the objective of inferring the polarity of a tweet. For this experiment, only positive and negative *tweets* are considered. This is due to a couple of reasons, the first one being that neutral and irrelevant *tweets* take the most part of the corpus (almost 80%).

TABLE XV
CLASSIFICATION METRICS COMPARED FOR DIFFERENT MACHINE LEARNING TECHNIQUES. BOLD CELLS INDICATE THE BEST VALUE FOR EACH METRIC

	ACC	PPV	NPV	TPR	TNR	MCC
OpinAIS	73.31%	69.52%	79.03%	**83.33%**	63.23%	47.54%
Naive Bayes	70.42%	83.33%	64.65%	51.28%	89.68%	44.33%
C4.5	66.88%	70.87%	64.13%	57.69%	76.13%	34.40%
Random Forest	**79.42%**	78.75%	**80.13%**	80.77%	78.08%	**58.86%**
SVM	65.59%	83.56%	60.08%	39.10%	92.26%	36.99%
MLP	71.06%	**85.87%**	64.84%	50.64%	91.61%	46.29%
CLONALG	63.99%	68.64%	61.14%	51.92%	76.13%	28.91%
AIRS	65.92%	81.25%	60.61%	41.66%	90.32%	36.59%
Immunos-81	63.99%	83.33%	58.78%	35.26%	**92.90%**	34.44%

While this handicap could be solved by balancing the classes, a second problem appears: words from the irrelevant *tweets* will take a considerable part of the features array, if not the whole. This happens given that irrelevant *tweets* contains many words in different languages, which only appear in *tweets* from that class and which turn out to be very good discriminators. For this reason, the classifier would specialize in neutral and irrelevant *tweets* rather than on positive or negative ones, which was the original problem.

The number of positive and negative *tweets* is balanced (47.83% vs. 52.17% respectively). The confusion matrix for this problem is shown in Table V, and Table VI shows different metrics for evaluating the prediction quality [13] for both the train as test sets, including the accuracy (ACC), the positive predictive value (PPV, also known as precision), the negative predictive value (NPV), the true positive rate (TPR, also known as recall), the true negative rate (TNR, also known as specify) and the Matthews correlation coefficient (MCC). Unclassified instances are treated as if they were misclassified. It can be seen that the results are fairly good and the generalization power is acceptable as well, as the performance over the test set is pretty similar to that over the training set.

Finally, a comparative evaluation is performed to check how the results obtained using OpinAIS for the task of sentiment analysis compare to those achievable using classic ML techniques. For this task, the Weka framework is used [15]. An ARFF file is generated directly from the OpinAIS framework, after the preprocessing and vectorization phases. Thus, the resulting ARFF is a file where instances have 200 binary attributes (one per each selected word) and a class, either *positive* or *negative*.

For the comparative evaluation, classic ML techniques have been used (Naive Bayes, C4.5 decision trees and random forest), as well as Kernel-based techniques (support vector machines) and biologically-inspired algorithms (multilayer perceptron). Also, alternative immune classifiers[4] have been compared, namely CLONALG [9, 10], AIRS [43] and Immunos-81 [4]. In all cases, default parameters are used.

Confusion matrices displaying the classification results for each of this techniques are shown as follows: Table VII shows the confusion matrix using Naive Bayes, Table VIII for C4.5 decision trees, Table IX for random forest, Table X for support vector machines (SVM), Table XI for multilayer perceptron (MLP), Table XII for CLONALG, Table XIII for AIRS and Table XIV for Immunos-81.

All these results are synthesized in Table XV where the metrics for evaluating the prediction quality described above are computed for the results obtained in the test set for each technique. Results show that for class-independent metrics (accuracy and Matthew's correlation coefficient), OpinAIS performs better than most of the other classifiers, with the only exception of random forests. In a per-class basis, OpinAIS provides the best results for the true positive rate (TPR), i.e., is able to classify most positive *tweets* correctly, outperforming the other techniques. On the other hand, the results are worse for the true negative rate (TNR), meaning that many negative *tweets* are either misclassified or not classified at all. Also, the negative predictive value (NPV) compares quite well to the alternative techniques, only surpassed by random forest, meaning that most of the *tweets* predicted as negative are really negative. From all the techniques compared, OpinAIS is the second with highest generalization power (measured as the absolute difference between the accuracy of the training and the test set), outperforming all its competitors expect for Immunos-81.

VII. CONCLUSIONS AND FUTURE WORK

As a result of the present work, a framework for applying IS to a variety of classification problems, including those involving sentiment analysis or some natural language processing, has been developed. While this framework is initially built as an implementation of an evolutionary algorithm, it has been refactored to keep extensibility as the main priority. This way, scientists can easily adapt the framework to their needs, either adding new algorithms or information retrieval processes or supporting new input data.

The first framework prototype has been evaluated by using two different public datasets. Results are encouraging, as binary classification metrics for the evaluated datasets are always greater than 50% and in some cases close to 100% and MCC is significantly higher than zero, and the built classifiers proved to generalize fairly well the concepts they learnt. When these results are compared to other machine learning techniques, OpinAIS outperforms them in terms of accuracy, with the only exception of the random forest classifier, and in any case behaves significantly better in a class-independent basis than its immune-based competitors.

[4] Added to WEKA as a plugin available at: http://wekaclassalgos.sourceforge.net

As the framework remains in a phase of active development, many improvements can be proposed as future work. For instance, many new algorithms, parameters and information retrieval features can still be added, and n-grams rather than words could be used for vectorization.

APPENDIX

A. Running the Framework

The OpinAIS class, which is placed in the root package (es.uc3m.baldo.opinais), contains the entry point for the application. This class only requires an argument, which is the path for a *properties* file (a special type of file in Java, very similar to *.ini* files), containing a bunch of parameters.

CODE I
A SAMPLE PROPERTIES FILE

```
# Set of possible types (classes).
types=apple,google,microsoft,twitter

# Reader which will retrieve the items from the input source.
reader=TweetReader

# Factory which will process and convert the items to individuals.
factory=TextIndividualsFactory

# Source file with input.
inputFile=data/SandersAnalytics/tweets_brand.txt

# Maximum number of individuals. 0 means all.
individualsSize=0

# Must the number of individuals for each type be balanced?
isBalanced=false

# Size of the population of detectors.
speciesSize=200

# Length of the features vector.
featuresLength=1000

# Preprocessors to be applied, in order.
preprocessors=LowerCaser,StopWordsRemover,Stemmer

# Percentage of the individuals to be used in the test set.
testPct=0.1

# Name of the algorithm to be used.
algorithm=EvolutionaryAlgorithm

# Types of the arguments required by the constructor
algorithmTypes=Integer,Double,Double,Double,Double,Double
```

A fragment of a sample properties file is shown in Code I. Besides, the distributed source code contains also properties files for some applications, which the user may want to take a look at to get a better understanding of all the parameters that can be customized.

B. Extending the Framework

The purpose of this section is to provide a brief overview to developers on how they can extend the framework to support new inputs.

For this example, the *US Congressional Voting Records Data Set* [36] described before has been chosen. This dataset contains a set of instances representing a certain congressman, which can be either republican or democrat. Each of these stores the particular vote of the congressman for 16 different votations, where this vote can be a *yes*, a *no*, or an abstention.

The steps for supporting classification over this dataset are the next ones:

1) In the first place, the developer must create a class in the *ir.items* package, which represents a vote record and may be called *VotingRecord*. This class must extend from *Item* and will store the vote for each votation.
2) Secondly, a class converting input lines into instances of the *VotingRecords* class will be implemented, and stored in the *ir.readers.factories* package, while implementing the *Factory* interface. This class may be called *VotingRecordFactory*.
3) Later, a reader in the *ir.readers* class will be developed, which must implement the *Reader* interface. This reader will eventually return a set of voting records given an input file.
4) A vectorizer must be developed to encode voting records as a binary string. To do so, the approach of [41] can be observed, where yes is represented as *01*, no is represented as *10* and abstention is represented as *00* (notice that this encoding is not arbitrary, and it has been chosen so that opposite values differs in its genomic representation as much as possible). This class can be called *VotingRecordVectorizer* and must be placed in the *ir.vectorizers* package.
5) Finally, the individuals factory implementing the interface *IndividualsFactory* must be developed, which essentially coordinates the flow between the classes above to generate a set of individuals.

ACKNOWLEDGMENT

This work was partially funded by the Spanish Ministry of Science and Innovation under MOVES project (TIN2011-28336) and European Union's CIP Programme (ICT-PSP-2012) under grant agreement no. 325146 (SEACW project).

REFERENCES

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment Analysis of Twitter Data. In Proc. LSM, pages 30 - 38, 2011.
[2] E. Cambria, T. Mazzocco, and A. Hussain. Application of Multi-Dimensional Scaling and Artificial Neural Networks for Biologically Inspired Opinion Mining. BICA, 4:41-53, 2013.
[3] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New Avenues in Opinion Mining and Sentiment Analysis. IEEE Intelligent Systems, 28(2):15-21, 2013.
[4] J. H. Carter. The Immune System as a Model for Classification and Pattern Recognition. Genetic Programming and Evolvable Machines, 7:28-41, 2000.
[5] H. Chen and D. Zimbra. AI and Opinion Mining. IEEE Intelligent Systems, 25(3):74-76, 2010.
[6] H. Cordobés, A. Fernández-Anta, L. F. Chiroque, F. Pérez, T. Redondo, and A. Santos. Graph-based Techniques for Topic Classification of Tweets in Spanish. IJIMAI, 2:31-37, 2014.

[7] S. Dasgupta and V. Ng. Mine the Easy, Classify the Hard: a Semi-Supervised Approach to Automatic Sentiment Classification. In Proc. ACL-IJCNLP, pages 701-709, 2009.
[8] L. N. de Castro and J. Timmis. Artificial Immune Systems: A New Computational Intelligence Approach. Springer, 2002.
[9] L. N. de Castro and F. J. Von Zuben. The Clonal Selection Algorithm with Engineering Applications. In Proc. GECCO, pages 36-37, 2000.
[10] L. N. de Castro and F. J. Von Zuben. Learning and Optimization Using the Clonal Selection Principle. IEEE Transactions on Evolutionary Computation, 6:239-251, 2002.
[11] L. Dey and S. M. Haque. Opinion Mining from Noisy Text Data. IJDAR, 12:205-226, 2009.
[12] K. D. Elgert. Immunology - Understanding the Immune System. John Wiley & Sons, Inc., 1996.
[13] T. Fawcett. An Introduction to ROC Analysis. Pattern Recognition Letters, 27:861-874, 2006.
[14] M. Govindarajan. Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm. IJACR, 3(4):139-145, 2013.
[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11(1):10-18, 2009.
[16] R. R. S. Jandail, P. Sharma, and C. Agrawal. A Survey on Sentiment Analysis and Opinion Mining: A Need for an Organization and Requirement of a Customer. IJETAE, 4(3):17-24, 2014.
[17] A. N. Jebaseeli and E. Kirubakaran. Genetic Optimized Neural Network Algorithm to Improve Classification Accuracy for Opinion Mining of M-Learning Reviews. IJETTCS, 2(3):345-349, 2013.
[18] J. Khairnar and M. Kinikar. Machine Learning Algorithms for Opinion Mining and Sentiment Classification. IJSRP, 3(6), 2013.
[19] E. Kouloumpis, T. Wilson, and J. Moore. Twitter sentiment analysis: The good the bad and the omg! In Proc. ICWSM, 2011.
[20] S. Leon-Rojas, U. Kirschenmann, and M. Wolpers. We Have No Feelings, We Have Emoticons ;-). In Proc. ICALT, pages 642-646, 2012.
[21] S. Li, Z. Wang, G. Zhou, and S. Y. M. Lee. Semi-Supervised Learning for Imbalanced Sentiment Classification. In Proc. IJCAI, pages 1826-1831, 2011.
[22] B. Liu. Sentiment analysis: A multifaceted problem. IEEE Intelligent Systems, 25(3):76-80, 2010.
[23] B. Liu and L. Zhang. A Survey of Opinion Mining and Sentiment Analysis. In Mining Text Data, pages 415-463. Springer, 2012.
[24] D. Maynard, K. Bontcheva, and D. Rout. Challenges in developing opinion mining tools for social media. In Proc. LREC Workshop, pages 15-22, 2012.
[25] N. Mishra and C. K. Jha. Classification of Opinion Mining Techniques. IJCA, 56(13):1-6, 2012.
[26] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In Proc. LREC, pages 1320-1326, 2010.
[27] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. FTIR, 2(1-2):1-135, 2008.
[28] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs Up? Sentiment Classification using Machine Learning Techniques. In Proc. EMNLP, pages 79-86, 2002.
[29] M. F. Porter. An Algorithm for Suffix Stripping. Program, 14(3):130-137, 1980.
[30] M. A. Potter and K. A. De Jong. The Coevolution of Antibodies for Concept Learning. In Proc. PPSN, pages 530-539, 1998.
[31] M. Puteh, N. Isa, S. Puteh, and N. A. Redzuan. Sentiment Mining of Malay Newspaper (SAMNews) Using Artificial Immune System. In Proc. WCE, 2013.
[32] A. Puurula and S.-H. Myaeng. Integrated Instanceand Class-based Generative Modeling for Text Classification. In Proc. ADCS, pages 66-73, 2013.
[33] Y. Saez, C. Navarro, A. Mochón, and P. Isasi. A System for Personality and Happiness Detection. IJIMAI, 2:7-15, 2014.
[34] N. Samsudin, M. Puteh, A. R. Hamdan, and M. Z. A. Nazri. Is Artificial Immune System Suitable for Opinion Mining. In Proc. DMO, pages 131-136, 2012.
[35] N. Samsudin, M. Puteh, A. R. Hamdan, and M. Z. A. Nazri. Immune Based Feature Selection for Opinion Mining. In Proc. WCE, 2013.
[36] J. C. Schlimmer. Concept Acquisition through Representational Adjustment. PhD thesis, University of California, Irvine, 1987.
[37] B. Schuller and T. Knaup. Learning and Knowledge-Based Sentiment Analysis in Movie Review Key Excerpts. In Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues, volume 6456 of LNCS, pages 448-472. Springer, 2011.
[38] N. R. Sharma and V. D. Chitre. Opinion Mining, Analysis and its Challenges. IJIACS, 3(1):59-65, 2014.
[39] G. Stylios, C. D. Katsis, and D. Christodoulakis. Using Bio-inspired intelligence for Web opinion Mining. IJCA, 87(5):36-43, 2014.
[40] S. Tan, X. Cheng, Y. Wang, and H. Xu. Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. In Advances in Information Retrieval, volume 5478 of LNCS, pages 337-349. Springer, 2009.
[41] J. Twycross and S. Cayzer. An Immune-based Approach to Document Classification. Technical report, HP Laboratories, Filton Road, Stoke Gifford, Bristol U.K., 2002.
[42] G. Vinodhini and R. M. Chandrasekaran. Sentiment Analysis and Opinion Mining: A Survey. IJARCSSE, 2(6), 2012.
[43] A. Watkins, J. Timmis, and L. Boggess. Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm. Genetic Programming and Evolvable Machines, 5:291-317, 2004.
[44] J. Wilbur and K. Sirotkin. The Automatic Identification of Stop Words. JIS, 18:45-55, 1992.

Alejandro Baldominos Gómez (M'14) was born in Madrid (Spain) in 1990. He obtained his bachelor in computer science and engineering from Universidad Carlos III de Madrid (Spain) in 2012 and his master in computer science and technology in 2013 with a specialization in artificial intelligence from the same university. He has been working as a Research Assistant at Universidad Carlos III de Madrid in the Advanced Databases Group since 2010, and in the Evolutionary Computation and Artificial Intelligence Group since 2013. He is now a Ph.D. student with a studentship granted by the Spanish Ministry of Education, Culture and Sport. He also works as Professor of the master in visual analytics and big data at Universidad Internacional de la Rioja. He has published several conference papers in the fields of context-aware systems, artificial intelligence and big data and has been involved in several national and European research projects, such as Semants, Cadooh, Memento and SEACW. Prof. Baldominos is AAAI Student Member and IEEE Student Member.

Nerea Luis Mingueza was born in Madrid (Spain) in 1991. She obtained her bachelor in computer science and engineering from Universidad Carlos III de Madrid (Spain) in 2013 and her master in computer science and technology in 2014 with a specialization in artificial intelligence from the same university. She has held a research fellowship since 2012 in the Planning and Learning Group of Universidad Carlos III de Madrid, where she is now a Ph.D. student. She is also Professor of the bachelor in computer science and engineering, where she teaches the programming course. Her research fields are humanoid robotics and multi-agent planning, having several papers of these topics published in conference proceedings.

Mª Cristina García del Pozo was born in Madrid (Spain) in 1983. She obtained her computer management engineering degree in 2006, her master in computer security in 2007 and her industrial organization degree in 2009, all of them from Universidad Pontificia de Salamanca. She is currently studying the master programme in technological projects design and management from Universidad Internacional de la Rioja. She has participated in several national and European projects, including Yuste Digital and GBIC at Universidad Pontificia de Salamanca, and SEACW and E-Space at Universidad Carlos III de Madrid, where she currently works as Research Assistant in the Evolutionary Computation and Artificial Intelligence Group. She had previously worked as JAE-Tec in CSIC (Consejo Superior de Investigaciones Científicas) for two years developing projects, including the platform used for managing researchers' normalized curriculum vitae (CVN).

Security Framework for Agent-Based Cloud Computing

Venkateshwaran K, Anu Malviya, Utkarsha Dikshit, S.Venkatesan

Department of Information Technology, Indian Institute of Information Technology Allahabad

***Abstract* — Agent can play a key role in bringing suitable cloud services to the customer based on their requirements. In agent based cloud computing, agent does negotiation, coordination, cooperation and collaboration on behalf of the customer to make the decisions in efficient manner. However the agent based cloud computing have some security issues like (a.) addition of malicious agent in the cloud environment which could demolish the process by attacking other agents, (b.) denial of service by creating flooding attacks on other involved agents. (c.) Some of the exceptions in the agent interaction protocol such as *Not-Understood* and *Cancel_Meta* protocol can be misused and may lead to terminating the connection of all the other agents participating in the negotiating services. Also, this paper proposes algorithms to solve these issues to ensure that there will be no intervention of any malicious activities during the agent interaction.**

***Keywords* — Agents, Cloud Computing, Security, Contract Net Protocol, Service Capability Table, Agent Trust Table**

I. INTRODUCTION

CLOUD computing is a fast developing technology which provides scalable data storage to large and various services without the hassle of installation and maintenance. Since there is an increase in number of users for the cloud services, there is a demand on cloud service providers. Hence there is a need for dynamic and automated cloud service composition [1, 2].

With the emergence of large number of service providers, users are not able to choose the best cloud service based on their technical and financial requirements.

To address this problem agents are introduced in the cloud environment. These agents make the decision making process easier for consumers by choosing and providing the best fit service for them, based on their requirements. According to Kwang Sim's model of agent based cloud composition [1], there are four agents involved in the cloud commerce such as Consumer Agent (CA), Broker Agent (BA), Service Provider Agent (SPA) and Resource Agent (RA).

Every agent will maintain a SCT (Service Capability Table), the attributes of SCTs are: (i) agents' addresses (ii) the requirements that agents can resolve, and (iii) the last known status of the service [1]. The SCT gets updated with agent's status after each and every agent-agent interaction.

All the agents interact with each other to deliver the best cloud service to the user. The whole process of agent interaction is controlled by the semi recursive contract net protocol (SR-CNP). This protocol is used to do the negotiation process between the task managers and the contractors. Here, the agent who initiates the process and requires a task to be done is referred as task manager and the agent who is able to execute the task is known as contractor.

There are various interaction protocols that can be followed for the agent-agent interaction. The model of agent based cloud commerce requires recursive call for, proposal and acceptance at various stages. So the contract net protocol in a semi recursive manner suits well into the model.

There are two roles in the contract net protocol: (i) Initiator (ii) Participant [6].

A consumer adopting the initiator role broadcasts a call-for-proposals to achieve a task (e.g., service composition) to n participants (contractors). The participants may reply with: (i) a proposal (quotation) to carry out the task, or (ii) a refuse message.

From the received m proposals, the initiator will select the best (cheapest) proposal, and sends: (i) an accept-proposal message to the best participant, and (ii) reject-proposal messages to the remaining (m – 1) contractors [1].

After carrying out the task, the selected participant sends either: (i) an inform-result message or (ii) a failure message in case of unsuccessful results.

This paper briefs about the mechanism of agent based cloud computing in section II, explains the security issues with agent based cloud computing in section III, and proposes solutions to overcome the issues in section IV and conclusion in Section V.

II. MECHANISM OF AGENT BASED CLOUD COMPUTING

A. Consumer Agent

Consumer agents receive requirements from the consumers. Each consumer agent maintains an SCT, which contains list of several broker agents known to it. Whenever a user requests for a cloud service, consumer agents receive these requests and sends a call for proposal to all the broker agents with a certain timeout, say 30 seconds, to respond.

Broker agent responds with either accept or reject based on its ability to resolve the request. Consumer agent only accepts those responses which come within the timeout period, other responses are discarded.

Among all the responses received, CA selects the most suitable BA, sends the accept-proposal to the selected BA and

DOI: 10.9781/ijimai.2015.334

refuse-proposal to all other BAs.

B. Broker Agent

Broker agents provide a single virtual cloud service to the consumers by contacting and selecting set of Service Provider Agents (SPA). BAs act as an intermediate between the CAs and SPAs.
Every BA has two SCTs:

a) List of SPAs

b) List of other BAs (to be used in case of sub contracts required).

BAs also handle the update requests from the consumer agent. BA selected by the consumer agent selects the SPAs from its list, makes a contract with SPA and delivers service to the consumer agent.

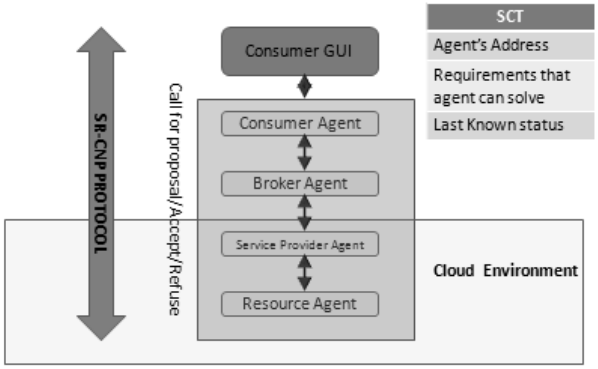


Fig 1. Mechanism of Agent-Based Cloud Computing

C. Service Provider Agent

On the agreement of transactions, SPA allocates and de-allocates the cloud resources from the resource agents.
Every SPA has two SCTs.

a) List of Resource agents (RA)

b) List of other SPAs for the subcontracts.

SPAs keep track of the available resources and synchronize with the RAs for concurrent or parallel executions. Selected SPA approaches the available RAs and makes the contract for the consumer requirements.

D. Resource Agent

Resource agents are the major control agents for accessing cloud resources. RAs are associated with SCT table consisting of SPAs. Whenever there is a request from SPA, RA sends resource or status to SPA based on the availability of resources.

As depicted in Figure 1,Once RA sends resource to SPA, resources are delivered to BA, BA delivers the cloud service to CA and consumer gets its service from CA.

Agents use predefined built-in functions [1] for sending messages to other agents.

The process is bounded by the timeouts. It involves two timeouts, timeout1 and timeout2. Timeout1 refers to the deadline of proposal submission and timeout2 refers to the deadline to deliver the virtual service.

This mechanism, delivers the cloud service to the consumer by making use of agents.

III. SECURITY ISSUES IN AGENT BASED CLOUD COMPUTING

Agent based cloud computing is developed in an ideal environment. Agents have been introduced to mainly focus on the process of negotiation for choosing the best cloud resource for the consumer. Since agents are the third parties, there are lots of security issues involved. This paper identifies several security issues which can block the agents from choosing suitable resources.

1. Addition of Malicious agent

Unlike the acquaintance network which updates the agent list only during the addition of new agent, SCTs update the agent list whenever a transaction happens between the agents. Though this feature of SCTs improves the performance of message exchange and always keeps the updated information about agent in the table, there is a security threat in addition of new agents.

According to Kwang's model [1], SCTs can add a new agent into the list when there has been a previous encounter with the agent or by mere presence of an agent in the same cloud.

In this scenario, any malicious agent can add itself into an SCT and can receive all the consumer requirements associated with it.

Following are the possible impacts when a malicious agent gets added into an SCT.

a) Getting involved in all consumer requirement negotiations thus misguiding the process by providing unrealistically cheap prices and blocking other legitimate cloud resources from providing services to the consumer.

b) Capturing the responses of other agents and sending spoofed messages to the initiator and other participant agents.

2. Flooding Attack

To keep the records updated, SCTs get updated with agent's status whenever a transaction between agents occurs.For Example consider, a broker agent sends a call for proposal request to all the SPAs given in the SCT. Suppose the broker agent's SCT contains a malicious agent then during the broadcast of call for proposal for a consumer requirement, malicious agent gets a message. Now, the malicious agent can flood the response to the initiator agent (i.e. Broker agent) with its response as accept the proposal and status of the agent as available.

Until the timeout, initiator agent will receive all the responses from SPAs and update its SCT. When a malicious agent creates flooding response, SCT will be involved in updating the information of the malicious agent only.

3. Exceptions to Protocol flow

FIPA has mandated few exceptions in the agent interaction which should be present in every multi-agent system to control the flow of the process. Some of such exceptions like Cancel_Meta protocol and Not-Understood problem can be misused by the intruder agents.

These exceptions can be used for attacks as explained below:

(i) Forced termination of agent interaction

As per FIPA interaction protocol flow [6], any interaction between agents is identified using a globally unique and non-null conversation-id parameter. In a multi-agent environment having no security measure, a malicious agent can get involved in some other agent-agent interaction.

A broker agent sends a call for proposal to all the SPAs in its SCT. Agents reply with accept/reject messages. Any malicious agent can send a spoofed message with conversation-id and agent's address stating that context of the message is not-understood.

Not-Understood is a communicative act in the FIPA so that an agent should be able to handle errors when the semantics followed by different agents are different. When any agent does not understand the context of message sent by the sender, then the receiver can send a Not-Understood message, in this case sender will handle the error and terminate the connection with receiver. This can be exploited by malicious agent because on receiving the Not-Understood message, Broker agent terminates the connection with the SPA.

Further, response of the legitimate SPA will be discarded by the broker agent. Hence, there will be a forced termination of the connection between agents.

(ii) Artificial timeout creation

When an agent sends any request to other agent, it receives the response within the timeout period. As per FIPA exception of protocol flow, there is a provision that a sender can cancel the previously sent request by sending a Cancel_Meta protocol message to the receiver. On receiving Cancel_Meta protocol message, receiver thinks that sender no longer requires response for the request sent.

In cloud environment, any malicious agent with conversation-id and agent address can send a Cancel_Meta protocol. On receiving the message, receiver ignores the request sent from the sender while sender is still waiting for the response from receiver until the timeout period. Hence the artificial timeout created by malicious agent stops the receiver agent from sending the response to sender agent.

IV. PROPOSED SOLUTION

Proposed framework consists of various modules includes Security Agent, serving as front end authenticator and trust analyzer of a cloud. Other sections depict solution for various identified security issues.

Security Agent

Among the four agents (CA, BA, SPA, RA) involved in cloud computing, SPA and RA are created by the respective clouds and are called as cloud agents. Remaining CA and BA are referred as outsider agents. These outsider agents especially BAs interact with SPAs to get a cloud's service.

Hence, entry of malicious agent may occurs when BA come to interact with an SPA to request and negotiate for a requirement. So, a new entity known as Security Agent (SA) is introduced for every cloud environment (Fig. 2) to handle the outsider agents.

SA provides two services:

i. Verification

ii. Trust Degree Analysis

To interact with SPA and RA of a cloud environment, an outsider agent should be authenticated by the Security Agent every time (Fig.1).

(i) Verification

When an agent comes to interact with any cloud environment, SA should verify the agent with Agent Trust Table. If the agent record is not available in ATT, it is considered as New Agent to the cloud environment. The agent details will be added to ATT after verification process from Third Party. If the agent record is available in ATT, agent is already registered by SA and considered as Registered Agent.

For Registered Agent, SA should check for agent's authentication on its proxy server with the credentials. If the authentication process fails the agent is discarded with no more further processing. If the authentication is successful, the agent is allowed to interact with the cloud agent and then the trust degree of the replying agent will be analyzed and updated in ATT.

Thus a secure environment for agent interaction can be created and this can resolve the addition of malicious agents into the cloud environment (Fig. 2).

(ii) Trust Degree Analysis

To maintain trust in agent interaction, a trust model can be used. When a trusted communication happens, the trust degree of the agent gets increase.

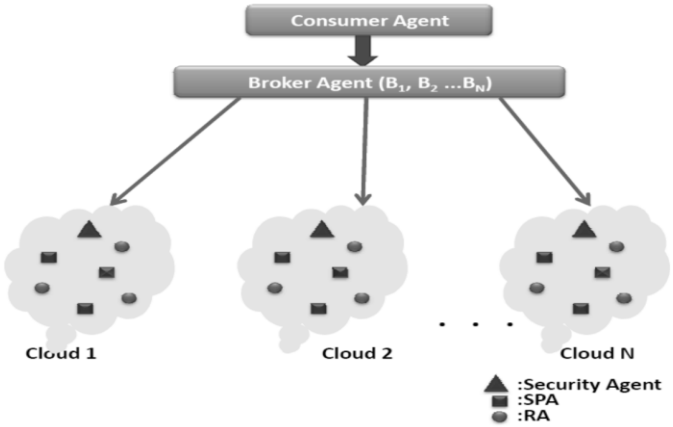


Fig 2. Agent Based Cloud Computing

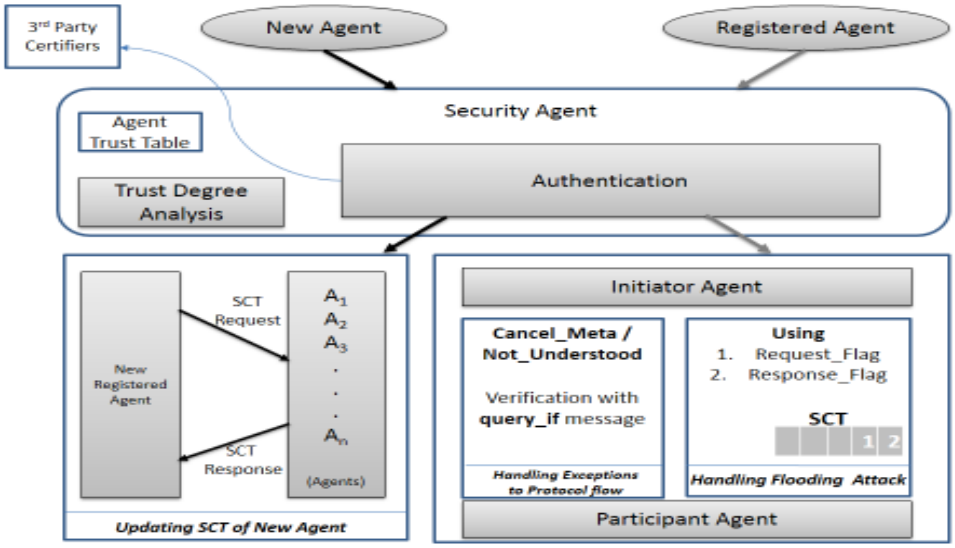


Fig 3. Framework for Secure Agent Communication

Similarly, when a non-trusted communication happens, the trust degree of the agent gets decrease. The probability of executing a request for any trusted agent is higher than the non-trusted or innocent agent. Suppose, there are n number of agents in a cloud with their Agent ID's = {AID1, AID2… AIDn}. If at any instance ith reply is analyzed for addition of its details in SCT table of New Agent (NA), then i Є {1, 2... n}. The Agent may be trusted, non-trusted or innocent.

Agent Trust Table (ATT)

A trusted agent's Trust Degree increases and decreases on completion of a process either successfully or unsuccessfully depending on its performance or set policies. Probability function is used to determine the trust degree of an agent replying with its SCT table. Based on the TD, agents will be marked as trusted, non-trusted or innocent. Actions for any task can be positive or negative.

There is a difference among the negative actions. It can be a wrong action or a malicious action. Positive actions are the right actions done by the trusted agent. Wrong actions are the bad actions that do not cause any damage or may cause damages done by the innocent agent and malicious. actions are harmful actions such as attacks done by the non-trusted agent.

The Trust Degree can be calculated by the equation: [4]

$$\textit{Trust Degree} \quad TD = \left(1 - \frac{N_a}{T_a}\right) A_w^{(s)} \quad \textit{Where } 0 \leq TD \leq 1$$

N_a = No. of negative actions
T_a = Total no of actions
A_w =weight of an action = 1 (for positive action)
0.9 (for negative action)
s = security level, s $\geq$ 1
Initially TD = 1; s =1
Threshold value =0.1

As the trust degree is calculated by exponential times of security level, if the positive action is happen with number of times (security level s =1,2…n), the term $A_w^{(s)}$ and should maintain the trust degree value . Hence, for positive actions, Aw is set to 1 and for negative actions, Aw is set to 0.9 to decrease the trust degree.

For example, suppose for a particular agent, Na is 50, Ta is 100 and the last updated behavior is positive and the s is 10th.

Then the Trust Degree from the above equation comes as TD = 0.5, which is greater than the threshold value i.e. the action is positive. Hence, its details will be added in the SCT.

Security Agent creates and maintains an Agent Trust Table (ATT) that include Agent's ID (AID), number of negative actions (Na) , total number of actions (Ta), Agents Behavior, security level (s) and action value or Trust Degree (TD) (Fig. 4).

Agent Trust Table				
Agent's ID (AID)	Number of negative actions (N_a)	Total number of actions (T_a)	Agents Behaviour	Security level (s)

Fig 4. Agent Trust Table.

It is used to check trustworthiness of either a newly created or previously registered agent through this calculated Trust Degree. The Agent Behavior of ATT is used to account the action weight of that agent depending upon its behavior either positive or negative. This ATT is updated every time after completion of a transaction.

On every completion of a transaction, this Trust Degree is calculated and ATT is updated with latest Agent Behavior by the Security Agent.

Addition of New Agent

Security Agent (SA) authenticates new agent arriving in cloud and handles Trust Degree for updating Service Capability Table (SCT) of that new agent.

When a New_Agent (NA) arrives in cloud environment, it goes to Security Agent (SA) which authenticates this NA by checking its Agent_id (AID) in trusted third part, if present,

Algorithm for Addition of New Agent:

Input: New Agent
Output: Addition or Discarding of New Agent

1 New_Agent (NA) arrives in cloud environment
2 NA goes to Security Agent (SA)
3 if SA (AID (NA) present in index of AMS)
4 Assign password for AID
5 Check Authentication on proxy server created by SA
6 if (AID (NA) && pwd == Correct)
7 Create SCT table
8 Broadcast SCT_details (AID)
9 if SCT_details (AID,reply (AID_i))
10 Send Request (AID_i ,TD) to SA
11 SA→ if (TD (AID_i) > threshold_val ?)
12 Update (ATT)
13 Update New_Agent (AID_i ,SCT)
14 Check for more Agent's reply Goto Step 9
15 else decrease (TD(AID_i))
16 Report AID_i action as negative to SA
17 if (no. of negative behaviour >= x)
18 Report AID_i as malicious to SA and Discard AID_i
19 else Goto Step 9
20 else Goto Step 8
21 else Discard NA

SA will assign a unique password. Here after authenticated by the cloud's proxy server created by Security Agent and SCT table of NA is created, otherwise this NA is discarded.

The NA now sends a broadcast message to all the other agents in the cloud environment to enter details into its SCT table. Since the issue was to avoid addition of malicious agent details, so the trustworthiness of the agent is measured for every ith reply coming with its SCT details. The NA requests SA to check the Trust Degree (TD) of ith agent, if it's greater than or equal to defined threshold value, the SCT detail of ith agent is updated in NA's SCT. If number of negative behavior identified by SA is greater than the threshold value, the reply is discarded.

This involves two processes:

(i) Authentication and
(ii) Trust Examining.

i. Authentication

Addition of new agent to the cloud environment: According to various research papers addition of agent can be based on

1. Trust: Where Certificate Authority (CA) serves as the root of trust.

CA issues these certificates only to those Principals who are trusted by the CA based on their harmless intentions and actions (Principal is a person who signs on behalf of the Agent code and is responsible for the behaviour of agent. Principal should be well aware of the workflow, behaviour and operational consequences of the agent).

2. Validation: When the owner registers the agent to the agent platform, this platform should validate the owner and log the request's source address.

Thus an agent arriving newly in a cloud environment must be signed for trust or be registered with a Third party who can guarantor for the Agent's behavior. This generates a unique identifier for each agent named as Agent Identifier (AID).

When a new agent (NA) wants to enter into a cloud, it reaches to Security Agent of that cloud which checks for its registration with Third Party by looking for its AID into their index, its Access permissions and its previous transaction or registration details with other clouds, to verify whether the coming agent is a legitimate agent or not. After verification if NA is found legitimate, SA assigns a password to it. The agent is now every time authenticated on cloud environment by its proxy server with this AID and password. Any discrimination from above checks leads to discarding of agent from interacting with cloud agents. The agent record is added in the Agent Trust Table (ATT) with default values. All the trusted agents of the cloud are added in the agent SCT broadcast list.

Algorithm for Solution of Forced Termination of Connection and Artificial Timeout:

Input: Reply from Particapant_Agent
Output: Accept or Discard the Reply

```
1  Initiator_Agent sends Call_for_Proposal
2  if Reply (Participant_Agent, Call_for_Proposal) == Accept
3       Connection (i) Initialisation
4       if Reply_i ( ) ==Cancel_Meta || Not_Understood
5        query_if (Reply_i, Reciever_Agent→Sender_Agent)
6          if query_if (Ack_i) ==True
7            Process Reply ( Cancel_Meta or Not_Understood)
8          else
          Ignore (Reply_i )
9  else
   Ignore (Reply ())
```

ii. Trust Examining

When a reply is received from an agent with its current SCT details, a request is sent to Security Agent (SA) with Agent_ID (AID) where current or updated Trust Degree (TD), present in Agent Trust Table (ATT) is checked or calculated and compared with Threshold Value. If the Agent's TD on an instance i is greater than the set Threshold Value, the ATT is updated and NA's SCT is updated with SCT details of replying agent (AIDi).

If current or updated TD is less than the threshold value, the action is said to be a negative action. TD of that agent (AIDi) is decreased as per the set policies and action is reported as negative or wrong to SA. If negative action occurs greater than or equal to x times, the action is reported as malicious and hence this replying participant agent (AIDi) is discarded from further processing.

The advantage is that the New_Agent (NA) remains unaffected when an identified participant agent does any malicious actions in the cloud environment. The Trust Degree of participant agent decreases accordingly with the malicious activities and the updating policies.

Handling flooding attack

To handle flooding attack issue, two flag attributes: *Request_Flag* and *Response_Flag* (*Fig. 5*), are introduced into the SCT along with the Agent's Address, Requirement provides and the Last Known Status. Since these are flags, so there values are either 0 or 1. Initially, both *Request_Flag* and

Response_Flag are 0. The values of these flags changes when a request is sent or a response is received for any agent.

Service Capability Table				
Agent's Address	Requirement provides	Last Known Status	Request_Flag	Response_Flag

Fig. 5. Service Capability Table

When an agent sends a message to other agent, represented as Request initiation, the Request_Flag in SCT sets to 1 and starts waiting for the response. As soon as the reply is recieved, it checks for the current status of flags i.e. if any response is received until now for that agent or not. If Request_Flag is equal to 1 and and Response_Flag is 0, the SCT of the Initiator_Agent is updated with Response_Flag as 1. Otherwise the response is discarded, showing that the reply from the agent for that request is already received. As soon as timeout occurs, the flag values are again set to 0.

Handling exceptions to Protocol flow

To avoid attacks on exceptions to Protocol flow the use of query_if function [5,10] is proposed. When Initiator_Agent broadcasts a Call_for_Proposal to all the other agents in the cloud, all the other agents reply either with accept or reject message depending on their willingness to communicate. The Initiator_Agent initializes the connection with all the agents replying as *Accept,* with unique Conversation_ID and *Reject* reply is ignored. If during the communication a *Not_Understood* or *Cancel_Meta* message is encountered, *query_if* function is initiated. The Receiver_Agent of these message sends a *query_if* message to the Sender_Agent and waits for the acknowledgment. If acknowledgment comes as *true*, the connection is terminated follow the message and act accordingly. Otherwise reply is ignored and the communication is continued.

Limitations and Implications

Though, we have mentioned that SA should refer to the trusted third party to verify the genuineness of an agent, it depends upon the cloud service providers to decide which trusted third party they want to believe. SA has to process each and every agent interaction occurs in a cloud. SA must be developed with the capability to handle maximum number of agents' queries at same time.

There may be possibility of discarding of an agent request, if the SA is not developed to handle multiple requests. However, this framework can be extended to determine the quality of the service offered by an agent. When the quality of the service can be compared, the user will get the most suitable service than the negotiation based on cost and time.

Algorithm for Flooding Attack

Input: Status Update Request from Participant_Agent
Output: Updating SCT or Discard the Update Request
SCT (Agent's Address, Requirement provides, Last Known Status, Request_Flag, Response_Flag)

```
Initially,
   Request_Flag = 0;
   Response_Flag = 0;
1   Initiator_Agent initiates Request ();
        Set Request_Flag = 1
2   receive Reply (Initiator_Agent, Participant_Agent)
3   if (Request_Flag == 1 && Response_Flag == 0) ==True
4        Update SCT ( , , ,1,1)
5   else
                  Discard Reply (Initiator_Agent, Participant_Agent)
6   if (Timeout)
7        Reset Request_Flag=0
           & Response_Flag=0
8   else
   Goto Step 2
```

V. CONCLUSION

Cloud computing is one of the futuristic technologies on which technology giants are counting. In future, number of users using the cloud computing is expected to increase gradually as there is a demand for cloud service exists. In such a scenario, there will no doubt that agents will play key role in selecting suitable services to users.

Since, it will be in the hands of agents to deliver a service to end user, agents should be free from attacks and bias. In this paper we have identified several security issues during the agent interaction. We have proposed solutions to handle those security issues. End User who uses the cloud services doesn't have any idea about how the agents are interacting and the service delivered is best among others or not. There is a possibility that malicious agent can involve in the process and

deliver wrong or malicious service to the user. So, we have used the trust degree analysis to decide whether the agent involved in the negotiation process is trusted or not. Analysis of this will help the proposed framework of security agent to allow only the trusted agents to deliver the service to end-user. However, several issues may arise when the agents plays dominant role such as determining the quality of the cloud service. With the proper security measures implemented in the cloud environment, agent based cloud computing will play as a platform for the consumers to use the perfect service.

REFERENCES

[1] Kwang Mong Sim, "Agent-based cloud service composition," Springer Science + Business Media, LLC2012

[2] Kwang Mong Sim, "Agent-based cloud commerce," Department of Information and Communications, Gwangju Institute of Science & Technology, South Korea

[3] Page, J.; Zaslavsky, A.; Indrawan, M., "Countering agent security vulnerabilities using an extended SENSE schema," Intelligent Agent Technology, 2004. (IAT 2004). Proceedings. IEEE/WIC/ACM International Conference on , vol., no., pp.183,189, 20-24 Sept. 2004
doi: 10.1109/IAT.2004.1342942

[4] Shantanu Pal, Sunirmal Khatua, Corresponding Author Nabendu Chaki, Sugata Sanyal, "A New Trusted and Collaborative Agent Based approach for Ensuring Cloud Security," in arxiv.org 2011
92 A. P. C. Road, University of Calcutta, India

[5] Fabio Bellifemine, Giovanni Caire,Dominic Greenwood in book "Developing Multi-Agent Systems with JADE,"
Telecom Italia, Italy.

[6] "FIPA Contract Net Interaction Protocol Specification,"
in http://www.fipa.org/specs/fipa00029/SC00029H.html

[7] Fatma Masmoudi, Monia Loulou,Ahmed Hadj Kacem, "Formal Security Framework For Agent Based Cloud Systems," in 2014 International Workshop on Advanced Information Systems for Enterprises ,ReDCAD Laboratory, University of Sfax, Tunisia.

[8] Suleiman Onimisi Aliyu and Kwang Mong Sim, IEEE Senior Member, "Minimizing Message Exchanges in Agent Based Cloud Service Composition," in Proceedings of the International MultiConference of Engineers and Computer Scientists 2014 Vol I,
IMECS 2014, March 12 - 14, 2014, Hong Kong

[9] Khushbu Virani, Dhara Virani, "Service Composition Based on Multi Agent in Cloud Environment," in International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 9, November- 2012

[10] "FIPA Request Interaction Protocol Specification,"
http://www.fipa.org/specs/fipa00026/SC00026H.htm

[11] Juan Pablo Paz Grau, Andrés Castillo Sanz, Rubén González Crespo "An Evaluation of Integration Technologies to Expose Agent Actions as Web Services," in Practical Applications of Intelligent Systems 259-270, January 1,2014, Springer Berlin Heidelberg

[12] Gutierrez, C., "An Analysis Architecture for Communications in Multi-agent Systems," International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI), Special Issue on Artificial Intelligence and Social Application Vol 2 Number 1 Pagination 65-72, March, 2013

[13] Jordán Pascual Espada, Vicente García Díaz, Rubén González Crespo, Oscar Sanjuán Martínez, B Cristina Pelayo G-Bustelo, Juan Manuel Cueva Lovelle, "Using extended web technologies to develop Bluetooth multi-platform mobile applications for interact with smart things," Information Fusion Vol 21 Pages 30-41,January 31,2015, Elsevier

S.Venkatesan received his Ph.D in Computer Science and Engineering from Anna University,Chennai, Tamil Nadu, India. He is an Assistant Professor in Department of Information Technology at Indian Institute of Information Technology, Allahabad. He has published various papers in numerous international conferences and journals. He is an active member in Association for Computing Machinery (ACM) Cryptology Research Society of India (CRSI). His current research includes Mobile Agent Security, Cryptography and Cloud Computing and Social Network Privacy.

Anu Malviya is a MS (Cyber Law and Information Security) research student in Indian Institute of Information Technology, Allahabad. She completed her Bachelor of Technology in Computer Science from Uttar Pradesh Technical University Lucknow. She is an active IEEE member. She has served as Trainer-Information Security in Mphasis, Bangalore as an intern. Her current research area includes cloud computing, IT Governance, Risk & Compliance.

Venkateshwaran K received his B.Tech in Information and Technology from Anna University, Coimbatore, Tamil Nadu, India. He is pursuing his Master of Science in Cyber Law and Information Security at Indian Institute of Information Technology, Allahabad. He worked as Software Engineer for two years in the domain of Mainframe Systems. His current research area includes cloud computing, Data Protection and Privacy, Risk assessment.

Utkarsha Dikshit is a MS (Cyber Law and Information Security) research student in Indian Institute of Information Technology, Allahabad. She has completed her Bachelor of Technology from SRM University, Chennai.

The Use of Genetic Algorithms in UV Disinfection of Drinking Water

Hugo R. Zaldaña[1], Emerson Castañeda[2]

[1]College of Engineering, Pontifical University of Salamanca, Spain
[2]Computer Science Faculty, Universidad Complutense de Madrid, Spain

***Abstract* — In order to have drinking water, some countries have to use chlorine. It is use cause is effective and it's cheap. An alternative to this process is the UV disinfection of drinking water. Most of the devices in the market use UV bulbs or mercury lamps. The UV LED, which is cheaper and smaller, allows creating new smaller devices. The main contribution of this paper is the use of Genetic Algorithms to help design a drinking water device with UV LEDs.**

***Keywords* — Genetic Algorithm, UV light, Water disinfection, Ultraviolet disinfection.**

I. INTRODUCTION

IN the last years, we can find many devices that use the Ultra Violet disinfection technologies. It has been used in many different devices: Systems for the control of airborne pathogens and allergens in indoor environments [1], pond pumps and soap for phones or even Reptile Bulbs for taking care reptile pets. There is also portable UV water purifier like the SteriPEN Sidewinder, or even the wastewater system from trojanUV. Some uses mercury lamps, UV bulbs, and some started to use UV LEDs. One of the reasons of the use of LEDs is because nowadays are cheaper and last longer.

But just because LEDs are cheap, does not mean that we should put a lot of LEDs inside a box to disinfect water. We can optimize how many LEDs use, searching for the minimum LEDs to purify water. The genetics Algorithms are good for optimization problem solving. We can use them find how many LEDs are the minimal to disinfect water in a given space.

Genetic algorithms have boomed in recent decades. Its basic principles were proposed by Holland [2], and are well described in several books, as Goldberg [3], Poli, Langdon, McPhee [4], Michelle [5] or Davis [6]. They had been used for Traffic Signal Timings Optimization [7], for transport management in docks [8], to improve the Production of Biochemical Pathway [9], for recognizing the position of Ezafe construction in Persian Texts [10], or even for GPGPU Implementation of a Genetic Algorithm for Stereo Refinement [11]. Just to name a few. On this paper we will use them to design a disinfection box of drinking water with them.

The goal of the genetic algorithm is to find a set of parameters that minimizes the output of a function. [12]. GA differs from most optimization methods, because they have the following characteristics [3].

1) They work with a coding of parameters, not the parameters themselves.
2) They search from many points instead of a single point
3) They don´t use derivatives
4) They use random transition rules, not deterministic rules.

This paper is structured as follows: Section II is a brief overview of the UV Light water disinfection. Also, a brief overview of LED's. Section III explains the implementation of a Genetic Algorithm to design a water purifying device using UV LED light. In section IV some results are presented and in section V some conclusions are drawn.

II. ULTRAVIOLET DISINFECTION

a. Brief history of Ultraviolet disinfection

The ability of sunlight to destroy microbial life has been known since 1877, with Downes and Blunt who reported that bacteria were inactivated by sunlight, and found that the violet-blue spectrum was the most effective [13]. This is caused by the invisible ultraviolet (UV) rays. With the development of the mercury – vapor lamp in 1901, the UV radiation technology was established. The low cost of chlorine for disinfecting of water slowed the development of UV-radiation technology, until 1950. In that year, the use of this technology began to increase in Europe. [14]

b. Disinfection Theory

Ultraviolet germicidal irradiation (UVGI) is defined as electromagnetic radiation in the range of about 200 to 320 nm that is used to destroy microorganisms. UVGI systems for water and surface disinfection have demonstrated reliability and effectiveness for the past century [1].

The term "germicidal" implies that these UV systems destroy, kill, or inactivate microorganisms. Microorganisms such as fungi, bacteria and viruses. The germicidal effectiveness peaks at about 260-265 nm. This peak corresponds to the peak of UV absorption by bacterial DNA [13]. The next table summarizes the definitions of the primary bands of UV radiation.

DOI: 10.9781/ijimai.2015.335

Dr. Kowalski writes that the definitions of the UV bands UVA, UVBA and UVC given in the table, have not yet been fully incorporated into every relevant guideline or adopted by every agency, but likely to be adopted eventually and universally. [13] In fact, you can buy UVC germicidal light bulbs or LED's asking UV-C or UV-B.

TABLE I
PRIMARY BANDS OF ULTRAVIOLET RADIATION [13]

Band	Wavelength (nm)	Type and classification
VUV	100-200	Vacuum ultraviolet. Actinic.
UVC	200-280	Ozone producing. Germicidal. Actinic.
UVB	280-320	Erythema. Germicidal. Actinic
UVA	320-400	Non germicidal. Near-UV, Black light

.

The design of UV systems for water disinfection differs from surface and air disinfecting systems. The reason is that the attenuation of UV irradiance in water occurs within about 15ms and this necessitates both closely packed arrays of UV lamps and higher UV power levels. That is the reason that water disinfection required ten times higher doses that air disinfection applications.

c. Advantages of ultraviolet radiation over other systems

1) Secure destruction of all microorganisms.
2) Do not alter or modify the chemical composition of water, or its smell or taste.
3) There is no possibility of formation of harmful secondary compounds.
4) No corrosive action on the system.
5) No corrosive action on the system. Simple, clean and safe handling without risks or discomforts for maintenance.
6) Easy installation. [15].

d. Exposure time

Like all the disinfectants, the exposure time is vital to ensure good performance. This depends on the type of flow; as well as the characteristics of the system. However, the period must be related to the dosage required. Normal exposures are of the order of 10-20 seconds [16].

e. Brief introduction to how LED's work

LED stands for Light Emitting Diode. The first commercial LED was offered by the Texas Instruments Corporation in the early 1960s. The LED emitted infrared radiation near 870 nm. The manufacturing quantities of the product were low. Probably caused by the high price for one single LED, which reportedly was 130 US$ [17]. Today, you can buy a single LED for 0.01 to 0.50 US$ in www.alibaba.com or www.amazon.com

LEDs are semiconductors devices which gives off light when powered. The color is determined by the properties of the semiconductor material. The semiconductor material is in the form of a small chip. This sits in a small reflector which together with the case shape, determines the overall shape of the emerging beam.

LEDs have 2 connections: The Anode and the Cathode. The Anode is the positive connection and the Cathode is the ground connection (negative connection).

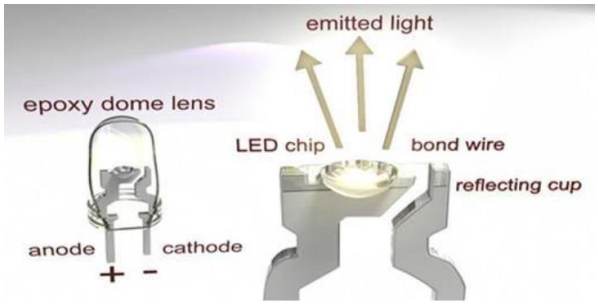


Fig. 1. Inside a LED, where we can see the reflecting cup [17]

f. Types of LED

There are many different sizes and shapes of LED's. They can be grouped based on their surface: planar, hemispherical and parabolic (fig 2). The most common are show in the fig 3. There are also High power LED's.

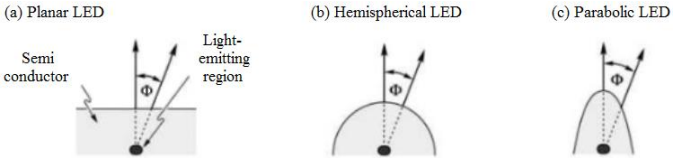
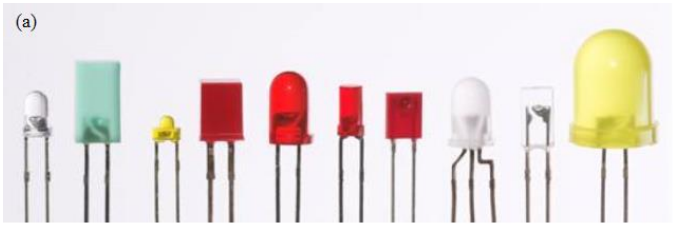


Fig. 2. Light emitting diodes with planar (a), hemispherical (b) and (c) Parabolic surfaces [16].



RGB LED (Red, Green, Blue) and Yellow

High power LED's

Fig. 3. (a) Common LED's on the market. (b) High Power LED's [17]

As we said, the overall shape of the emerging beam is determined by the small reflector together with the case shape. The Far-filed patterns of the different type of LEDs are show

in the fig 4. We won't talk about here about the lambertian emission pattern, just show that an angle of φ = 60°, the lambertian emission pattern decreases to 50% of its maximum value occurring at φ = 0|. The three emission patterns are normalized to unity intensity at φ = 0°, as show in fig 4.

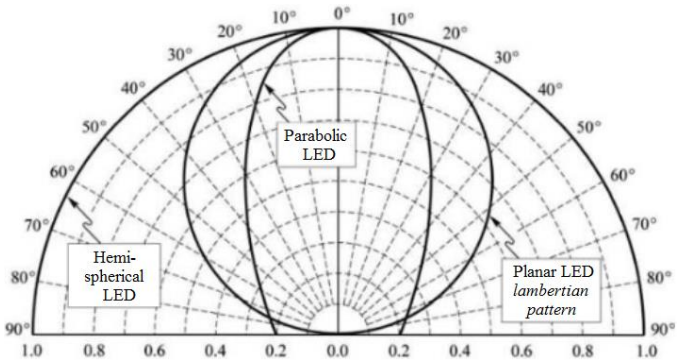


Fig 4. The Far-filed patterns of the different type of LEDs [16].

III. Design of devices using Genetics Algorithms

a. Antennas and GA

One of the most famous cases of the use of genetic algorithms, is the creation and patenting of an antenna, which was designed using a Genetic Algorithm [4] Genetic algorithms have been successfully applied to the design of a variety of antennas. [19]. We can name the design of thinned array in 1994 by Randy Haupt [12], wire antennas by Derek Linden in 1997, and Derek Linden with Altshuler in 1998 and 1999 [20], patch antennas by M. Johnson and Rahmat-Samii in 1999, linear and planar array by D. Marcano and F. Duran in 1999 [21]. The NASA also work in Evolutionary Algorithms to design Evolutionary Antenna [22].

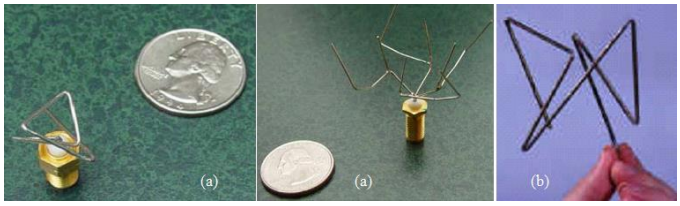


Fig. 5. (a) Evolutionary Antennas by NASA [21] [4]. http://www.nasa.gov/centers/ames/news/releases/2004/antenna/antenna.html. (b) Linden Antenna. http://www.lindenir.com/2.html

A book called "*Electromagnetic Optimization by Genetic Algorithms"* describes numerous applications of the genetic algorithm to antenna design [21]. But there is no research done for the use of Genetic Algorithms in UV disinfection of drinking water.

b. Disinfection of drinking water and GA

The same way that genetic algorithms are used for the antenna design, the principles of a genetic algorithm are going to be applied to help design a water purifier using UV LEDs. This is for academic purposes, for further construction and validation of the prototype.

The algorithm will seek to optimize the use of LED light in the device, in order to minimize the LEDs to be used. This will reduce consumption. To do this, we define that water will be deposited in a cube of 15 cm^3, inside the device that will purify the water. We establish that water comes without solid pieces that can protect viruses and bacteria from the germicidal effect of UV light. We do this with a filter placed at the inlet of water.

c. Chromosomes

We established that the cube is 15 cm per side, as the binary representation of 15 is "1111". Then, for each point with 3 coordinates (X, Y, Z) requires 12 genes. These coordinates indicate where the LED is located. Also we add a gene for LED orientation. "0" if the LED is down pointing up. "1 if it is up pointing down. Therefore, our minimum chromosome set as follows: (1)

$$DXXXXYYYYZZZZ = > 0101000110111 \quad (1)$$

This example will translate into up direction (0), with XYZ= 10, 3, 7 (1010 is 10, 0011 is 3, 0111 is 7). If we need to add another LED in the chromosome, 13 more genes will be added. With this we establish a proper representation of the chromosome.

d. Fitness

First, for this genetic algorithm, we will establish that the light is emitted from a Flat Top Led UV, and which forms a sphere of light from the LED, fig 6.

Fig. 6. (a) Flat Top Led Light Lamp LED UV

To define which chromosome is more fit than another, we will check how many points are left uncovered. This can be calculated by establishing that each centimeter of each axis (X, Y, Z), we have a small cube of 1 x 1 x 1 cm^3, which it will have a value of 1. We have 16 possible positions for each axis (0 to 15), obtaining a total of 4,096 positions to be illuminated by the LEDs inside the cube.

Additionally, we know that the attenuation of the UV light in water occurs about 15 cm [13]. Therefore, we establish that for this AG the diameter of the sphere of light is 15 cm, and its radius is 7.5 cm. The genetic algorithm will seek how many LEDs and where they should be placed so that the cube is fully illuminated by the spheres.

We count the number of genes that has the chromosome that is being evaluated. Then divided by the number of genes representing the coordinate of an LED (13). This way we get how many LEDs are on that chromosome.

For each LED, the position and orientation is obtained by the genes. We create a sphere from this point, and will travel up or down, depending on their orientation. Each point within the sphere is changed to 0. This can be done walking the Z axis, from 0 to 15. A chromosome that leave a space of 50 unlit cells is better than other one that leaves chromosome 75.

To define the points of the sphere that will be reached by the UV light, we start from the spherical coordinates, and switched to the Cartesian coordinates. Spherical coordinates are represented by the radius r, and θ and φ angles. These elements relate to the Cartesian coordinates using the same origin, as shown in Fig. 7

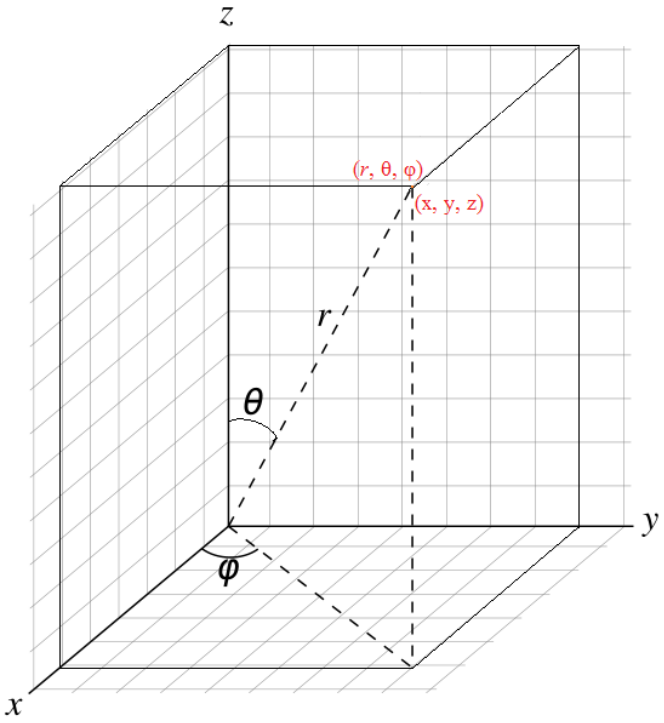


Fig. 7. Spherical coordinates and Cartesian coordinates.

The coordinates (X, Y, Z) of the point (r, φ and θ) are obtained based on trigonometric functions.

$$x = r \operatorname{sen} \theta \cos \varphi$$
$$y = r \operatorname{sen} \theta \operatorname{sen} \varphi \qquad (2)$$
$$z = r \cos \varphi$$

Now, we put the value of φ = 0 °. Therefore Cosine 0 ° = 1. Sine 0 ° = 0. Thus, X is 0, Y is simplified, and Z does not change.

$$x = r\,(0) \cos \varphi \Rightarrow x = 0$$
$$y = r\,(1) \operatorname{sen} \varphi \Rightarrow y = r \operatorname{sen} \varphi \qquad (3)$$
$$z = r \cos \varphi \Rightarrow z = r \cos \varphi$$

By having the angle φ at 0 ° and X = 0, we have two dimensions, as shown in Fig. 8

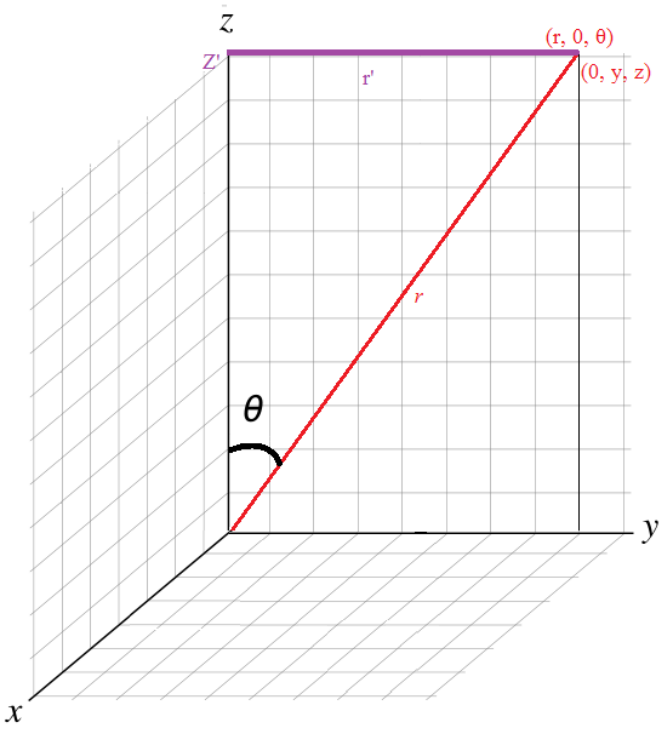


Fig. 8. 3D change to 2D

If we see the sphere from above, we can see several circles of different circumferences. *r'* is the radius of the circle resulting from the sphere see it from above for that level of Z. As we know the value of z ', which is the axis that the AG will walk along, and r is the radius of the sphere. If we keep z' and *r* constant, the angle θ is kept constant. We know the value of r and Z '. Now, we can calculate the value of r' with the Pythagorean theorem, where we know the hypotenuse, because is the radius of the sphere, and Z'. (4)

$$r^2 = Z'^2 + r'^2 \qquad r' = \sqrt{r^2 - Z'^2} \qquad (4)$$

Setting the value of r 'as constant, we have a cylinder. Using cylindrical coordinates we obtain X and Y as follows:

x = r' cos φ
y = r' sen φ
z = h, where h = z'

Then, changing first Z (-7 to 7, steps 1), and then changing the angle φ 1-360 °, we can have the points in the sphere, moving the origins to add the coordinates of the LED. Then, in the process, to explore the angle of 1° to 360°, we get the value of X and Y. Then, we keep the value of X as a constant, and Y axis is traversed, from "-Y" to "Y" to place a value of 0 at that coordinate. For these calculations, discard decimals, approaching to the lower integer. Also, the radius we will use will not be 7.5, will approach the lower integer, so we'll use 7.

The AG must stop when the sum of all cells in the cube is 0. In other words, all the cells are illuminated.

e. Mutation

We use the standard mutation, with a little change. If the GA not obtain a solution after n generations, for the mutation, we added 14 more genes. That is, add another LED.

IV. RESULTS

With the GA completed, we have results:

- It took on average 85.7 generations.
- 10 runs,
- Maximum of 132 generations and a minimum of 46 generations
- Maximum size of chromosome solution of 169 and a minimum size of 117 chromosome solution

Chromosome:
1101110110100100111100101110011010010010110011100111110100010011001110010011000001100110100001001011011101001001011010

This means 9 LEDs (13 gens * 9 = 117).
The coordinates of the LEDs can be read in Table II.

TABLE II
PRIMARY BANDS OF ULTRAVIOLET RADIATION [13]

Direction	X	Y	Z
Up	11	11	11
Up	3	12	12
Up	3	4	4
Down	12	14	14
Up	10	2	2
Down	14	4	4
Down	3	3	3
Down	4	11	11
Down	9	5	5

V. CONCLUSIONS

In this paper, an implementation of a genetic algorithm has been proposed to help design a disinfection device of drinking water using UV Light LED.

Thanks to the Genetic Algorithm, we can simplify the behavior of the UV light in water. Using the attenuation of UV light in water and using a planar LED. We can adjust the GA with other LEDs and knowing the shape of the emerging beam. We can change the space we are going to use it: Could be a big pipe in a Wastewater Treatment Facility.

For future research, it is the construction of the device, and add features that were not taken into account with the genetic algorithm to the water purifier. The genetic algorithm was useful to establish what is the minimum number of LEDs to disinfect water, in a given space.

Additions features are power supply of the LEDs, and resistance are not taken into account in the algorithm. The power supply could come from solar cells, or a dynamo, to use it when needed, even if there is no light.

Additionally, the design may include UV light sensors, and place them in the coordinate where the sphere ends. Thus an alert if needed would replace an LED.

For the work presented, we can conclude that it is feasible to apply genetic algorithms for water purification.

REFERENCES

[1] W. Kowalski. "Aerobiological Engineering Handbook. Airborne Disease and Control Technologies". McGraw-Hill, New York, ISBN 978-0-07-158882-9, (2006)

[2] J.H. Holland. "Adaptation in natural artificial systems". The MIT Press/Bradford Books, Massachusetts/ London, ISBN 978-0-262-08213-6 (1992), pp ix.

[3] D.E. Goldberg. "Genetic Algorithms in Search, Optimization and Machine Learning". Addison-Wesley Publishing Company, Inc., Massachusetts, United States of America, reprinted with corrections, ISBN 0-201-15767-5 (1989).

[4] R. Poli, W. B. Langdon, N. F. McPhee, "A field guide to Genetic Programming" (with contributions by John R. Koza), ISBN 978-1-4092-0073-4 (2008). Pp 119-120

[5] M. Mitchell. "An introduction to genetic algorithms". The MIT Press/Bradford Books, Massachusetts/ London, ISBN 0-262-13316-4 (1998).

[6] L. Davis (ed). "Handbook of genetic algorithms". Van Nostrand Reinhold, New York. ISBN 0-442-00173-8. (1991)

[7] A. A. Ezzat, H. A. Farouk, K. S. El-Kilany, A. F. Abdelmoneim. "Development of a Stochastic Genetic Algorithm for Traffic Signal Timings Optimization". Proceedings of the 2014 Industrial and Systems Engineering Research Conference, Y. Guan and H. Liao, eds. Montreal, Canada (2014).

[8] J. Villalba Montero. "Modelo de optimización por simulación para la gestión del transporte en muelles de la terminal de contenedores del puerto de Algeciras". Escuela Técnica Superior de Ingenieros de Sevilla, Universidad de Sevilla, Ingeniería Industrial, Proyecto de fin de carrera (2012) http://bibing.us.es/proyectos/abreproy/5073

[9] C. K. Chong, M. S. Mohamad, S. Deris, M. S. Shamsir, Y. W. Choon, L. E. Chai. "Improved Differential Evolution Algorithm for Parameter Estimation to Improve the Production of Biochemical Pathway". International Journal of Interactive Multimedia and Artificial Intelligence. Special Issue on Distributed Computing and Artificial Intelligence. (2012) http://www.ijimai.org/journal/node/223

[10] S. Noferesti, M. Shamsfard. "A Hybrid Algorithm for Recognizing the Position of Ezafe Constructions in Persian Texts". International Journal of Artificial Intelligence and Interactive Multimedia, Vol. 2, Nº 6 (2014) http://www.ijimai.org/journal/node/608

[11] A. Arranz, M. Alvar. "GPGPU Implementation of a Genetic Algorithm for Stereo Refinement". International Journal of Artificial Intelligence and Interactive Multimedia, Vol. 3, Nº 2. (2015). http://www.ijimai.org/journal/node/739

[12] R. L. Haupt. "Thinned Arrays Using Genetic Algorithms". IEEE Transactions on Antennas and Propagation, vol 42, No 7, July 1994. (1994), pp 993

[13] W. Kowalski. "Ultraviolet Germicidal Irradiation Handbook. UVGI for Air and Surface Disinfection", Kindle Edition, Springer Berlin Heidelberg; 1 edition (2010)

[14] L. Demers, R. C. Renner. "Alternative disinfection technologies for small drinking water systems". American Water Works Association. USA. ISBN 0-89867-635-5 (1992) pp 85 – 100.

[15] J. O. Santos Carrillo. "Sistema de control electrónico de PH y esterilización, de agua". Instituto Politécnico Nacional. Tesis de grado. México. (2009), pp. 15

[16] F. Solsona, J. P. Méndez. "Desinfección del agua", Centro Panamericano de Ingeniería Sanitaria y Ciencias del ambiente, Lima, Perú. (2002), pp 94.

[17] E. F. Schubert. "Light-Emitting Diodes". Second edition, Cambridge University Press, Cambridge, Kindle edition. ISBN 978-0-521-86538-8 (2006)

[18] D. McCrea. "Using LED's in your own projects, everything you need to know". Kindle Edition. Amazon Digital Services, Inc. (2014)

[19] J. R. Koza, M. A. Keane, M. J. Streeter, W. Mydlowec, J. Yu, G. Lanza. "Genetic Programming IV: Routine Human-Competitive Machine Intelligence". Kluwer Academic Publishers, USA, ISBN 1-4020-7446-8 (2003)
[20] D. S. Linden, E. E. Altshuler, "Evolving wire antennas using genetic algorithms: a review," Evolvable Hardware, 1999. Proceedings of the First NASA/DoD Workshop on, vol., no., pp.225, 232, (1999).
[21] Y. Rahmat-Samii (ed), E. Michielssen (ed). "Electromagnetic Optimization by Genetic Algorithms". Wily-Interscience. Wiley Series in Microware and Optical Engineering, Kai Chang, Series Editor. New York, ISBN 0-471-29545-0 (1999)
[22] G. S. Hornby, A. Globus, D. S. Linden, J. D. Lohn. "Automated Antenna Design with Evolutionary Algorithms". NASA.(2006)

Hugo Zaldaña was born in Guatemala in 1977. Currently is a project manager engineer and IT consultant in telecommunications business, as well a PhD Student at the Universidad Pontificia de Salamanca, campus Madrid (UPSAM). Currently, his areas of interest include Software Engineering, Genetic Algorithms, Genetic Programing, Water Disinfection, Math, and Artificial Intelligence.

Emerson Castaneda is currently an R&D engineer and IT consultant in the telecommunications business, as well a PhD Student at the Complutense University of Madrid (UCM). He received his Bachelor's Degree in Systems Engineering from Catholic University of Colombia (UCC) in 2001 and a MSc. (DEA) in Computer Science and Software Engineering from Pontifical University of Salamanca (UPSAM) in 2005. In 2007, he moved to the USA to study at the University of Texas and continued to the University of Denver (DU), where he was a Visiting Scholar. Castaneda has been a valued IEEE member for 12 years. Currently, his areas of interest include Real Time Analytics, Software Engineering, Patterns Matching, Soft Computing, Artificial Intelligence, and Computational Creativity.

Recognizing Human Activities Based on Wearable Inertial Measurements - Methods and Applications

Pekka Siirtola

Computer Science and Engineering Department, University of Oulu, Oulu, Finland

***Abstract* — On April 10 of 2015 Pekka Siirtola defended his PhD thesis, called "Recognizing Human Activities Based on Wearable Inertial Measurements - Methods and Applications" [1]. The thesis was supervised by Professor Juha Röning and pre-examined by Associate Professors Ulf Johansson from University of Borås, Sweden, and Daniel Roggen from University of Sussex, United Kingdom. Pekka Siirtola successfully defended his thesis against his opponent Professor Barbara Hammer from University of Bielefeld, Germany. This publicly open defence was held in Auditorium TS101 at University of Oulu, Finland.**

***Keywords* — Activity recognition, inertial sensors, pattern recognition**

I. Introduction

INERTIAL sensors are devices that measure movement, and therefore, when they are attached to a human body, they can be used to measure human movements. In the doctoral thesis "Recognizing Human Activities Based on Wearable Inertial Measurements - Methods and Applications" by Pekka Siirtola, data from these sensors are studied to recognize human activities user-independently. This is possible if the following two hypotheses are valid: firstly, as human movements are dissimilar between activities, also inertial sensor data between activities is so different that this data can be used to recognize activities. Secondly, while movements and inertial data are dissimilar between activities, they are so similar when different persons are performing the same activity that they can be recognized as the same activity. In the thesis, pattern recognition -based solutions are applied to inertial data to find these dissimilarities and similarities, and therefore, to build models to recognize activities user-independently.

The thesis is a compilation of six original publications ([2] – [7]) and it studies activity recognition in two contexts: daily activity recognition using mobile phones, and activity recognition in industrial context. Both of these contexts have special requirements and these are considered in the presented solutions. Mobile phones are optimal devices to measure daily activity: they include a wide range of useful sensors to detect activities, and people carry them with them most of the time. On the other hand, the usage of mobile phones in activity recognition includes several challenges; for instance, a person can carry a phone in any orientation, and there are hundreds of smartphone models, and each of them has specific hardware and software. In addition, in order to achieve highest possible recognition rates, all sensors need to be calibrated before they are used no matter if the used sensor is an inertial sensor or a camera [8]. Activity recognition using the inertial sensors of mobile phones are studied in original publications [2] and [3]. Moreover, as battery life is always as issue with smartphones, techniques to lighten the classification process are proposed (studied in original publications [4] and [5]). Industrial context is different from daily activity context: when daily activities are recognized, occasional misclassifications may disturb the user, but they do not cause any other type of harm. This is not the case when activities are recognized in industrial context and the purpose is to recognize if the assembly line worker has performed tasks correctly. In this case, false classifications may be much more harmful. Solutions to these challenges are presented in the thesis (studied in original publications [6] and [7]).

The solutions introduced in the thesis are applied to activity recognition data sets. However, as the basic idea of the activity recognition problem is the same as in many other pattern recognition procedures, most of the solutions can be applied to any pattern recognition problem, especially to ones where time series data is studied.

References

[1] Siirtola, Pekka (2015), "Recognizing human activities based on wearable inertial measurements: methods and applications", Department of Computer Science and Engineering, University of Oulu, April, Acta Universitatis Ouluensis C 524.

[2] Siirtola P & Röning J (2012), "Recognizing Human Activities User-independently on Smartphones Based on Accelerometer Data", International Journal of Interactive Multimedia and Artificial Intelligence (1), 5 (June 2012), 38-45.

[3] Siirtola P & Röning J (2013), "Ready-to-Use Activity Recognition for Smartphones". IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2013), 16-19 April 2013, 59—64.

[4] Siirtola P, Laurinen P, Röning J & Kinnunen H (2011), "Efficient Accelerometer-Based Swimming Exercise Tracking", IEEE Symposium on Computational Intelligence and D 15, 156--161.

DOI: 10.9781/ijimai.2015.336

[5] Siirtola P, Koskimäki H & Röning J (2011), “Periodic Quick Test for Classifying Long-term Activities”, IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2011), April 11-15, 135--140.
[6] Siirtola P, Koskimäki H, Laurinen P, Huikari V & Röning J (2011), “Improving the Classification Accuracy of Streaming Data Using SAX Similarity Features”, Pattern Recogn. Lett. 32, 13 (October 2011), 1659-1668.
[7] Koskimäki H, Huikari V, Siirtola P & Röning J (2013), “Behavior modelling in industrial assembly lines using a wrist-worn inertial measurement unit”, Journal of Ambient Intelligence and Humanized Computing: Volume 4, Issue 2 (2013), Page 187-194.
[8] Pacheco A, Bolivar-Baron H, Gonzalez-Crespo R, & Pascual-Espada J (2014), "Reconstruction of High Resolution 3D Objects from Incomplete Images and 3D Information", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 2, issue Regular Issue, no. 6, pp. 7-16.

Pekka Siirtola received his M.Sc. degree in Mathematics in February 2006 and Dr.Tech degree in April 2015, both from University of Oulu, Oulu, Finland. Currently, he is working as a post-doc at the Computer Science and Engineering Department at University of Oulu. His research interests include machine-learning aspects of signal shapes and, especially, activity recognition based on inertial measurements.

An Architecture Approach for 3D Render Distribution using Mobile Devices in Real Time

Holman Diego Bolívar[1], John Alexander Velandia[1], Jenny Natalia Torres[1], Elena Giménez de Ory[2]

[1]*Faculty of Engineering, Universidad Católica de Colombia, Bogotá D.C, Colombia*
[2]*School of Engineering, La Rioja, Spain*

***Abstract* — Nowadays, video games such as Massively Multiplayer Online Game (MMOG) have become cultural mediators. Mobile games contribute to a large number of downloads and potential benefits in the applications market. Although processing power of mobile devices increases the bandwidth transmission, a poor network connectivity may bottleneck Gaming as a Service (GaaS). In order to enhance performance in digital ecosystem, processing tasks are distributed among thin client devices and robust servers. This research is based on the method 'divide and rule', that is, volumetric surfaces are subdivided using a tree-KD of sequence of scenes in a game, so reducing the surface into small sets of points. Reconstruction efficiency is improved, because the search of data is performed in local and small regions. Processes are modeled through a finite set of states that are built using Hidden Markov Models with domains configured by heuristics. Six test that control the states of each heuristic, including the number of intervals are carried out to validate the proposed model. This validation concludes that the proposed model optimizes response frames per second, in a sequence of interactions.**

***Keywords* — Distribution render; Mobile devices; Hidden Markov models; Gaming as a service.**

I. INTRODUCTION

NOWADAYS, a vast network of recognized media, such as television, internet, game consoles, smartphones, tablets and desktop devices create new ways to play, to express oneself, learn, explore ideas and generate culture.

Computers are used as mediators in the learning process through play and social interaction. An example is the Massively Multiplayer Online Game (MMOG). In some scenarios, these games are considered an educational platform, because they allow players to learn together through personal interaction in a cooperative process. Recent studies reveal that, with the continued use of this type of games [1, 2, 3, 4], several learning processes are achieved (when creating a virtual identity, for instance).

Recent trends in mobile computing have truly commoditized a large number of components required for immersive virtual reality [3, 5]. Current thin client devices, such as smartphones and tablets, represent a renaissance in mobile computing. With gaming as a driver for the adoption of mobile graphics chipsets, these devices package unprecedented graphics dealing with position/orientation sensing, wireless networking, and high resolution displays. Such systems provide unique opportunities for constructing low-cost and mobile virtual reality systems [6].

Mobile games contribute to a huge number of downloads and, consequentially, to potential profits in the application market. However, although the processing power of mobile devices, as well as the transmission bandwidth is increasing, the unstable network connectivity may bottleneck the providing of Gaming as a Service (GaaS) for mobile devices. The hardware constraints of mobile devices, such as computational power, storage and battery, limit the representation of games [7].

Therefore, there is a need to reduce content and processing requirements, as well as to maintain control of storage and communication between users. For the mobile clients, one of the most problematic tasks is the presentation of the 3D Virtual Reality data. According to the 3D Virtual Reality data, the client has to calculate the position of objects, the lightning and shadows. This is a difficult task to perform with weak processors and a low main memory. In complex scenes, a high processing power is needed to process all data in nearly real time. A solution for this problem is the consumption of the processing task not on the client [8].

Render has significant features since there are a variety of methods to perform these virtual 3D graphics. In terms of software highlights there are four main algorithms: scanline rendering and rasterization, ray casting, ray tracing, and radiosity. Each of these algorithms is focused as a fundamental part in representing complex images, either by means of the light beam, or grouping pixels to reduce computer processing, or calculating the passage of light, etc. [9]. It must be taken into account that these processes also depend on the geometry applied in each algorithm.

Aside from the request processing with 3D content, it is also important to control the storage of data generated by applications and their communication in order to achieve decent interactivity between geographically dispersed users. Data are not centralized on a single server; therefore control is needed over scalability and fault tolerance to provide a

DOI: 10.9781/ijimai.2015.337

response to user requests [10]. Moreover, having a distributed system is an advantage for processing and user control.

A Platform as a Service (PaaS) allows virtualized computing resources via internet or advanced networks, allowing transparent use of resources. Along with offering storage services and computer processing, PaaS is built with Internet standards and protocols such as HTTP. A PaaS combines quality of service and broadcast functions distributed with capabilities in parallel processing. Together, these features create a platform for development software, designed specifically for network applications that produce and consume massive amounts of digital media. Thus, it is necessary to identify the technological architecture for gaming.

The reminder of this paper is organized as follows: section II summarizes the related work in industry and academy, and section III studies the process of subdivision surfaces to be held inside a mobile device. Section IV presents a model, including its architecture for distributed rendering based on hidden Markov model. The proposed model is assessed using performance tests according to frames per second. Section V presents the results of assessing the model, and section VI concludes our work, including future work to be developed from this research.

II. Related Works

Render 3D is used today to display molecular orbitals in the analysis of results of simulations of quantum chemistry [11], for dynamic medical evaluations, analysis of complex information models associated with medical training, management of geographically referenced information, and in the searching of extraterrestrial intelligence, among many other uses. [12]. In recent years, the performance and capacity of graphics processing units have improved dramatically, thanks to the parallelization of computational tasks [13], but an efficient operation of large capacities of parallelism, allowing a linear acceleration along with multiple compute nodes are still required [14], for they would optimize the graphics processing level data volume with polynomial complexity.

The display group NERSC and Lawrence Berkeley National Laboratory (LBNL) have developed the Visapult tool to attack these problems. Visapult is an application of parallel distributed processing that leverages the resources of computer networks and the processing power of supercomputers. Renders for volume ray tracing and traditional series can take many minutes or hours. Visapult supports interactive volume rendering to rates by employing distributed network components and a high degree of parallelism. Image Based Volume Rendering Algorithm when used with this program, Visapult can exchange additional information with reduced bandwidth [15].

For improving traditional visualization of render, Corcoran et al. [16] propose a model that employs two phases, which depend firstly on rendering volume direct (RVD), and on a number of other rendering non photorealistic processing techniques (RNF). By separating the visualization on two levels, allowing a higher level of detail than that normally observed with the traditional process, it is noteworthy its level software architecture. The interactivity drawback is due to the lack of specific limits and sometimes it is possible to get occlusion by overlapping images. Because being interactive, they require minimum time response.

The model poses strategies to solve these problems, emphasizing the perception of images.

Bounding volume hierarchies (BVHS) hold great promise for dynamic scenes. However, each proposed technique changes for handling animations has limitations, such as a reduced performance in a prolonged time and some difficulty in the processing of deformed objects. It avoids synchronization problems but in the other hand limits the speed at which BVH can build frames [17].

Madhavan et al., [18] show a model that seeks an implementation of a distributed rendering environment which is easily accessible according to the system requirements. The model generates the deployment work, with monitoring render, data sending, error corrections and reducing waiting times. Furthermore, Taura [19] proposes an architecture based on real time monitoring system called VGXP, based on a technology called GXP. For the system, it is important to monitor and control the performance of a distributed process, as well as the performance, scalability, fault tolerance, and also data sent to the client without overloading and security. The system generates a 3D graphics response in java.

In the system proposed with Kamoshida [20], the server collects the monitoring data required and sends it to the client through a hierarchical architecture. An agent process runs on each node, which monitors the control data produced by each process and event. To accomplish this communication, the agents form a tree structure for TCP connections. The root of the tree is the server process.

Madhavan et al. [21] propose a software architecture based on Java for real-time visualization and generating interactive graphics. This architecture minimizes the amount of required synchronization between PCs, resulting in excellent scalability.

The modular architecture provides a framework that can accommodate a variety of algorithms and data formats representation, provided that rendering algorithms are used to generate individual pixels and data duplication in each computer. As an object-oriented design, it implements the basic functionality required for distributed rendering.

Due to the complexity of volumetric rendering, the problem can be divided [22, 23, 24]. They propose using the Octree algorithm, which is responsible for dividing the volumetric scene in scenes less complex, according to the user's request. Another advantage is the weight of the scene at the time of transport on the network, since it does not require the entire bandwidth needed initially. The problem that arises with volumetric rendering is the volume size. Therefore, the image must submit to procedures outside the nucleus to avoid charges in memory.

Another algorithm commonly used for rendering is Ray-

tracing [15,16, 17, 25, 26, 27, 28]. This is based on the illumination of the image, capturing the beam size and its reflection on the object. This involves various drawbacks (shading, texture object, etc) when there are complex object.

In order to optimize performance, an intersection is found, the beam is transferred to the point of intersection and its address is modified according to the type of beam, shadow or reflection, using various characteristics of objects to be displayed and geometric data. The iterations stop when the reflected ray does not hit any objects or the maximum predetermined level of reflection is reached. Ray-traicing is a dynamic algorithm with a high cost in displaying images.

Castanie et al [29] propose a model based on an original application of DSM (Distributed Shared Memory), as it is a type of implementation to level hardware and software system. Each node in a cluster has access to a large shared memory that is added to the limited unshared memory of each node. However, this implementation is reconstructed, generating four additional access levels that are included in this system, such as the graphics memory, the local memory in the node, the memory of the other nodes through the network and the disk. This new implementation is called Distributed Hierarchical Cache System (DHCS).

III. Subdivision Surface Method on Mobile Device

The visualization process is composed of four parts: data collection, image processing, building surface and display of image. As regards techniques, there are two types of volume rendering methods: direct and indirect. Direct methods use a type of 3D volumetric images generated without explicitly extracting geometric surfaces from the data. Indirect method consists of marching cubes algorithm, from which the cells belonging to a surface threshold, and a threshold value provide as a result a cubic grid containing a classification of the object data, which is modeled through an octal tree.

Nowadays, the graphic processing is performed by the z-buffer algorithm which handles the display of images. It is useful because it processes millions of images interactively using triangles. It takes image texture and illuminates to a low computational cost. According to Shirley et al. [25], it has the following disadvantages:

- Applications with data sets significantly large, generates processing times of order NP.
- Applications with non-polygonal data are not easily converted into triangles for processing as image.
- Applications that demand high quality shadows, with reflection, refraction and particle effects are difficult to process.

An option is to perform the rendering process using ray-tracing techniques. However, a high computational cost is generated by the large number of ray tracing for each scene. This problem can be minimized by using special data structures able to organize or group scene objects spatially. The number of intersection tests involved in searching is greatly reduced.

So instead of following a comprehensive search to identify the correct scene for the nearest intersection, only nearby objects are approved and the remaining are discarded, as recommended by Siu-Lung et al. [26].

To develop volumetric render, tree octants Local Grade Smooth algorithm is used (OOLSD). It is submitted by Xing et al. [30]. First the image is divided into some small sets of points according to the octree construction, then a local triangular mesh through the region is built by fusing the triangulations recursively and aplying the principle of "divide and rule".

By reducing the surface into small sets of points, reconstruction efficiency is improved because the search region is small and local. In the recursive fusion process, the optimization operation is performed between boundary triangles. Therefore, the number of the mesh boundary triangulation does not increase, so the complexity of the algorithm is stable.

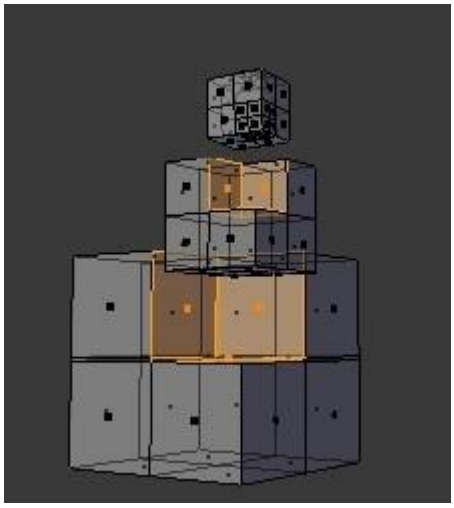


Fig. 1 Octree subdivision surface

Fig. 1 shows the subdivision surface from octants optimized algorithm, also known as Local Grade Smooth. The involved process of this algorithm reduces the level of memory which is stored in the mobile device by eliminating non-relevant images [22, 23].

For subdivision surfaces, it is necessary in the first place to identify those that are visible to users, through the hidden surface removal algorithm through JPCT 3D engine. Through CubMotion class structure renders is performed, based on the World Reference, FrameBuffer, Light, Object name, RGBColor Matrix, xPosition, YPosition, Object RenderOpenGL and URL ConstructiónXML. Given the description of the object in an XML file, the implementation is performed in the OpenGLRender class, which is responsible for the subdivision. By each tree node an event at BuildXML class is being created. BuildXML class is responsible for building tree-KD through spatial subdivision based on heuristics and surface areas. Taking into account the recommendation of Wald and Havran [31], Fig. 2 shows the

sequence diagram associated with the interaction surface subdivision process.

After the construction of the tree-KD, a session through the JSCH library is created, which is an implementation of the SSH2 protocol that provides support for secure remote access and data compression. In the SSH communication the following key issues are discussed:

- To access the server, it is necessary to provide credentials, user and password.
- Sockets are used to establish a communication between client and server. Thus, data transmission through objects is serialized.
- JSCH library provides an encrypted communication channel, protecting data that are travelling between client and server.

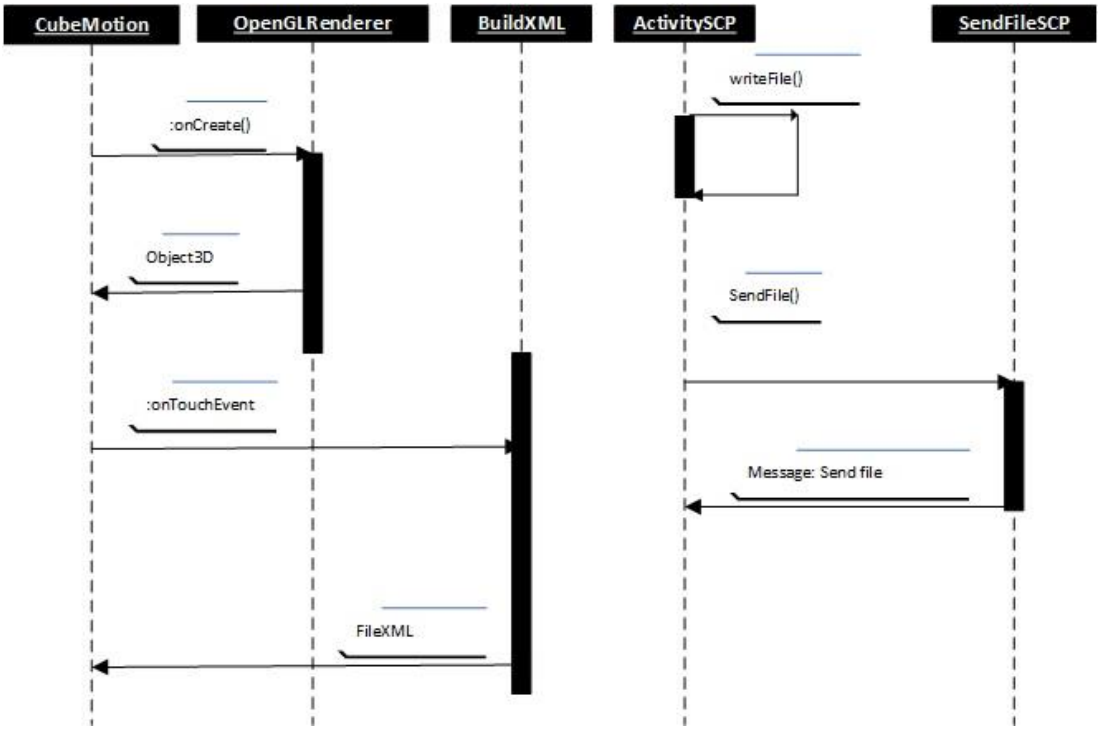


Fig. 2 Sequence diagram of surface subdivision

Once communication channel is created and a session is defined, the manner of handling events is throughout an implemented interface. ActivitySPC component is used to close sessions. Fig. 3 shows the deployment diagram associated with the process of subdivision surface on the mobile device and the creation of the communication channel.

Regarding the hardware environment used for testing, it encompassed a server and several mobile devices. The Server responsible for storing WMA has the following description: a blade server PowerEdge M620, Intel Xeon E5-2600 processor, Intel QuickPath Interconnect (QPI) 7.2 GT / s, 2.5 MB cache per core with 4 cores, 16GB RAM, 3TB HDD master was utilized and 3 TB hard drive slave, this server has a Matrox G200 video card integrated. This server has installed an operating system: Red Hat Enterprise Linux. Moreover, all test devices had the Android operating system, due to the need for the installation and configuration of JPCT 3D engine.

IV. Architecture Proposal for Distribution of Render

According to the tree-KD, metadata is generated for each node which is recorded in real time from a web service and relational database management system (RDBMS). This process consists of defining a structure with associated data that should map to an XML template in order to generate the BuildXML class. Fig. 4 shows the interaction diagram for each node.

For each mobile registered in the information system, a Workload Management Agent (WMA), similar to proposed in [20] is created. This agent is responsible for managing all render and sending requests to a scheduler that is in charge of being a matchmaker for compatible resources in a distributed system. Regarding the execution domain, it is based on Platform as a Service (PaaS). Inside this platform, there is a task called "Job Submission Service" responsible for sending tasks to the "Local Resource Manager". Then tasks are processed and a local scheduling is assigned to available resources schematized as worker nodes, which return the load to the WMA, which sends information mobile device.

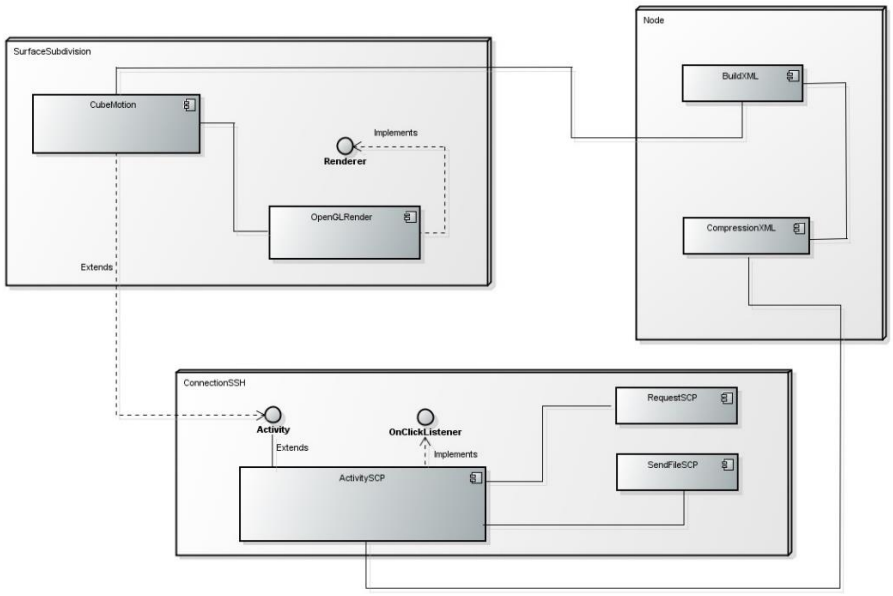


Fig. 3 Deployment diagram of surface subdivision

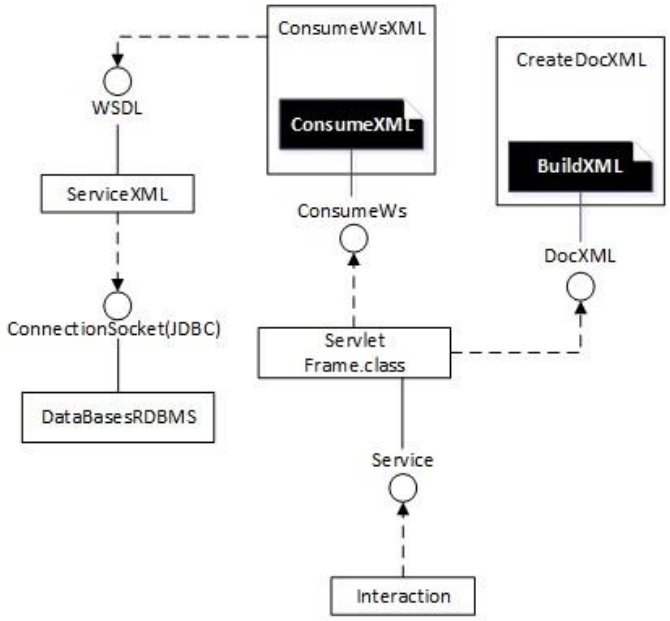


Fig. 4 Interaction diagram of nodes

Clearly, there is a dependence on bandwidth which is given by network congestion, the intensity of the received signal and the mobile device. Additionally, it should be considered that the following restrictions exist: network 2.5G technology and General Packet Radio Service (GPRS) transmitting 56 kbps. The aforementioned restriction inhibit data transmission to WMA since the standard for real-time animation is 24 frames per second (fps) with a limit of 292 Bytes for each XML document. Additionally, range between 1KB and 15KB acts as another restriction for data test.

Therefore, a minimum of 2880 kbps connection is required, which is only available in 4G networks.

One solution is to compress data using redundant coding bytes, through a grammar based on X3D standard. Then, decompression is performed in the WMA.

In the compression process, files are reduced to 500 B and 2.5 KB respectively, due to the redundancy of coordinates and XML tags.

Considering that the minimum value to establish a connection using compressed data is 480 kbps, it is still possible to perform tasks throughout 3G network. Regarding 2G networks with a top speed of 232 kbps, the rendering process requires applying a stroboscopic effect, reducing a third of the number of frames processed remotely.

To maintain the visual quality of animation in a game, mobile device should process two frames while the last frame is sent remotely. However, this simple model is not feasible due to ignorance of network traffic and the instability of the connection.

To identify the number of nodes created by each scene, we propose using Hidden Markov Models (HMM) that create a finite number of states, from an initial test of connection between the mobile device and WMA.

To model the full process, a known and finite set of states S Hidden Markov, one for each domain, is built. A domain is set from the combination of different heuristics such as:

- Connection speed.
- Latency time.
- Size of the scene.
- Number of nodes to construct feasible by scene.
- Probability of failure on the connection

Given that each of the established heuristics may vary in different real or integer values, prioritizing intervals according to initial conditions are established. One descendent prioritization is done according to the ideal conditions and the worst-case scenario and then it is assigned to each interval heuristic relevance considering the Hurwitz criterion. To calculate the weight of each interval heuristic $W[h(n)]$, in the

development of research, it was used:

$$W[h(n)] = \frac{W[h(1)]+W[h(2)]\ldots+W[h(n)]}{n} \quad (1)$$

$$W[h(n-1)] = W[h(n)] + \frac{W[h(1)]+W[h(2)]\ldots+W[h(n)]}{n-1} \quad (2)$$

In equation (1) weight or heuristic value is determined by assigning a weight to the last interval equal to the sum of the weights of the intervals over the number of intervals. Initial weight is 1 for each element. For the next element, previous element weight is taken and added to the result of the equation (2).

To calculate the relevance of each heuristic $R[h(i)]$, the value is set to 1 or 0 if the connection is in any of the established ranges and the importance, $I[h(i)]$, is indicated. The value obtained in (2) is multiplied by the value of the relevance of each heuristic, then added and this allows us to identify the importance of each heuristic. Equation (3), presents a formal method to identify the relevance of each heuristic.

$$R[h(n)] = \sum_{1}^{n} W[h(i)] * I[h(i)] \quad (3)$$

$R[h(n)]$ Sets the probability of heuristics associated, according to the connection status for each domain. A transition matrix between states is then generated.

In order to model the probability of the states, several vector observation, $V(i)$, have been established before the training phase models.

Formally, the probability to move from one state to another is represented by directed edges. Usually the nodes are numbered from 1 to N, according to the number of nodes and edges are labeled with probability values between 0 and 1. Each possible state $V(i)$, is represented by a labeled box and the probability that a job stay with some priority according to the policy node is expressed as a directed edge from that state to the observed symbol, as shown in Fig. 5.

Having defined the structure of the overall model and each of the HMM training proceeds of the S models to calculate the optimal values of all parameters that have been mentioned. For that, the k observation sequences in each state have been used.

For the proposed model the vector of initial probabilities A_i^s each HMM is initialized with probability value equal to 1 / N, as follows:

$$A_i^s = \frac{1}{N}, 1 \le j \le S \; and \; 1 \le i \le N \quad (4)$$

$B = (b_j(k))_N$ is a vector of observation symbol probability, one for each estate, in which $bj = (b_{j1}, b_{j2}, \ldots, b_{jM})$ defines observation symbol probability $ot = vk$ of the alphabet in the state j.

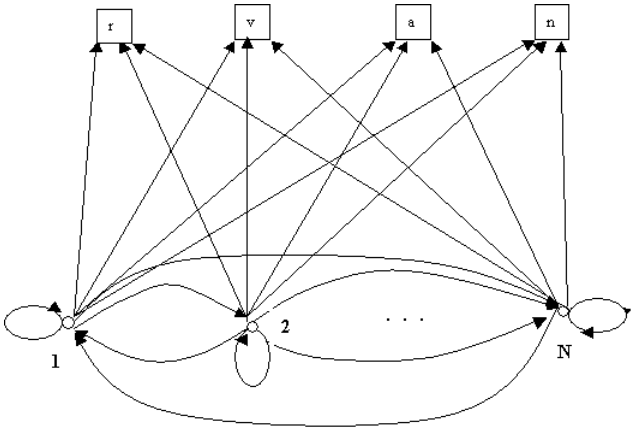


Fig. 5 Architecture fully connected (ergodic) model of Nodes

The learning model is to adjust the parameters to maximize P($V(i)$/λ) including several algorithms for training the HM Baum-Welch, Expectation Maximization (EM), Generalized Expectation Maximization (GEM), and different forms of gradient descent [30].

The procedure of training with EM for restimation of HMM parameters uses the variable $W[h(t(i,j))]$, that is the probability of being in state i at time t, and state j at time t + 1, for a given model λ and a observation sequence $V(i)$, i.e.

$$W[h(t(i,j))] = P[q_t = i, q_{t+1} \frac{j}{V(i)}, \lambda] \quad (5)$$

$$W[h\big(t(i,j)\big)] = \frac{P[q_t = i, q_{t+1} = j.V(i)/\lambda]}{P(\frac{V(i)}{\lambda})} \quad (6)$$

The sum of (5) on t can be interpreted as the expected number of transitions from state i to state j, formally expressed in (6):

$$\sum_{i=1}^{T-1}\sum_{j=1}^{N} W[h(t(i,j))] \quad (6)$$

Representing the expected number of transitions from state i to state j in $V(i)$. For optimal state sequence, the probability of the observation sequence $V(i)$, P($V(i)$|λ), is calculated efficiently.

The evaluation of the probability of the sequence consists of calculating the probability of the observation sequence P($V(i)$|λ). The way to solve the problem is to apply a forward algorithm. In this algorithm, it is assumed that αt (i) = P(o1 o2 ... ot, qt = i | λ). Then the probability of observing the partial sequence P($V(i)$|λ) in state i to time t can be calculated as follows,

$$\alpha_1(i) = A_i b_i\big(V(i)\big), 1 \le i \le N \quad (7)$$

$$\alpha_{t+1}(i) = \sum_{i=1}^{111t} \alpha_t(i) a_{ij}\ b_i\big(V(t+1)\big), 1 \le t \le T-1 \quad (8)$$

$$\mathrm{P}(V(i)|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \quad (9)$$

The observation probabilities are given by the state of the connection according to the values given by heuristic calculations taken into account (3).

WMA is responsible for maintaining a probability table for device and continuously monitors the connection status, acting as a brokered services under an implementation Gamming as a Service (GaaS) [4,5,7].

V. Results and Discussion

The render becomes a multi objective problem, since there is a problem of processing capacity on the mobile device.

The number of heuristics is increased, considering other factors such as device processing power, load balancing, cumulative yield, missed deadlines, equity, preference of users, total weighted completion time, delays weighted number of tardy jobs, and many others [33].

An animation render sequence was performed, starting at the scene shown in Fig. 6 and ending at the scene shown in Fig. 7. The animation corresponds to a 30 second walk of the main character (Fig. 7).

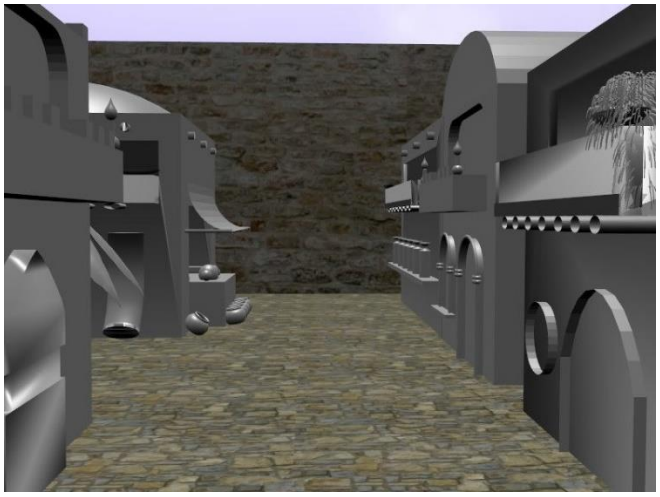


Fig. 6 Opening scene

The number of polygons of each scene varies between 7000 and 19000, with an average variation of 10% between frames.

In this research only the volume rendering process is taken into account, regardless of color, texture and lighting of the scene. For this reason buildings are in gray.

In order to validate the proposed model, six domains where considered and a controlled manner to stablish the states of each heuristic and the importance, $I[h(i)]$, associated with the number of intervals.

Table 1 shows the values of ranges for the best case, worst case and number of intervals associated to test domains according to heuristic set.

Fig. 7 End scene

Table 2 shows the domains where controlled tests were performed, which yielded the number of frames per second.

TABLE I
VALUES OF TEST DOMAINS USING HEURISTIC

	speed (Kbps)	*latency (ms)*	*Polygons (int)*	*Node-KD (int)*	*frames per second (int)*
Best case I[h(i)]	10240	100	7000	64	24
Worst case	128	250	19000	8	16
number of intervals	100	15	15	7	8

Table 3. Presents a variation of frames in mobile devices according to processor and memory. The conclusion is that the proposed model is valid for a set of significant number of polygons or a low-speed connection.

According to Table 3 and Fig. 8, there is not a significant variation between tests with domains 2 to 5, while in domain 1 there is a significant variation since it was considered the worst case in heuristics in Table 2.

Regarding the aforementioned, there is an inverse correlation between processor and memory capacity of the device, network speed and the number of frames per second if rendering process with JPCT 3D engine is performed. However, the correlation is reduced by 40% implementing the proposed model.

The rendering process is performed constructing a reference environment, commonly referred as world.

TABLE II
TEST DOMAINS

	speed (Kbps)	*latency (ms)*	*Polygons (int)*	*Node-KD (int)*
Domain 1	128	250	19000	8
Domain 2	10240	100	7000	64
Domain 3	1024	180	12000	32
Domain 4	2048	150	9000	64
Domain 5	512	100	15000	64

TABLE III
TEST RESULTS BY MOBILE DEVICE FEATURES

Mobile device				**Average frames per second**					
				Only JPCT 3D					**With WMA**
memory	**phone network**	**CPU**	**operating system version**	**Domain 1**	**Domain 2**	**Domain 3**	**Domain 4**	**Domain 5**	
2GB	HSPA+	Quad-core de 1,6 GHz Cortex-A15	Android V4.2 Jelly Bean	16,24	24,03	23,82	23,49	24,08	24,00
2GB	4G LTE	Quad-core 1.2 GHz Cortex-A53	Android 4.4 Kitkat	15,93	24,07	23,67	23,44	24,01	24,00
3GB	4G LTE	Quad-core 2.7 GHz Krait 450	Android 5.0 Lollipop	16,70	24,02	24,01	23,99	24,02	24,00
1GB	4G LTE	Quad-core 1.2 GHz Cortex-A7	Android 4.3 Jelly Bean	16,54	24,05	23,49	24,18	24,04	24,00
512MB	HSPA+	Dual Core 1.3 GHZ	Android V4.2 Jelly Bean	12,09	24,03	23,08	23,38	24,01	24,00
512MB	HSPA+	Dual-core de 1 GHz Cortex-A5	Android V4.2 Jelly Bean	12,50	24,03	23,04	22,88	23,98	24,00

Although reliability is tied to redundancy, this approach encompasses metrics of fault tolerance and system consistency. In that way, the expected reliability is covered, and the information system would cover the property of fault tolerance in its implementation. The proposed model was implemented and, during the phase of handling tasks, the rendering process was executed without interrupting the service.

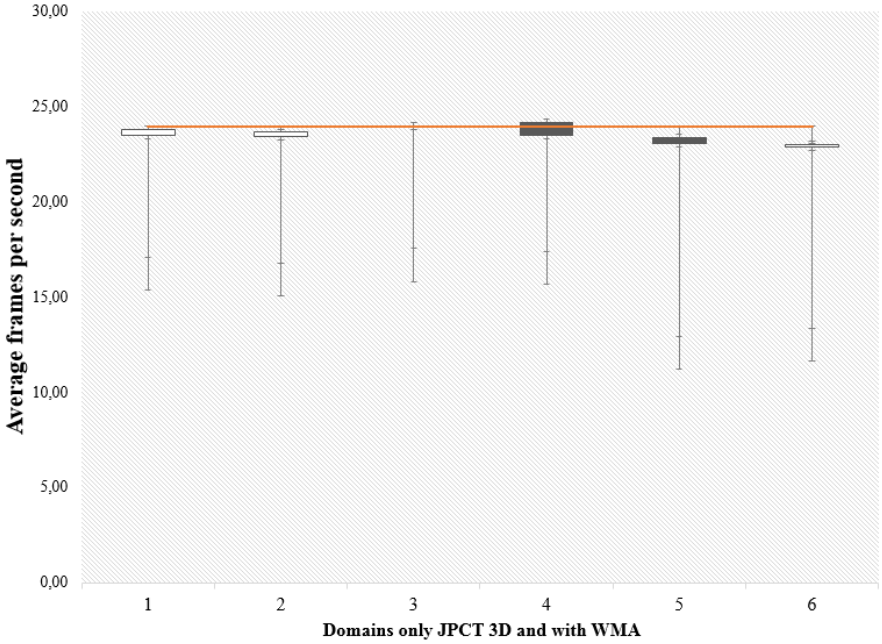


Fig. 8 Variation render between only JPCT and with WMA

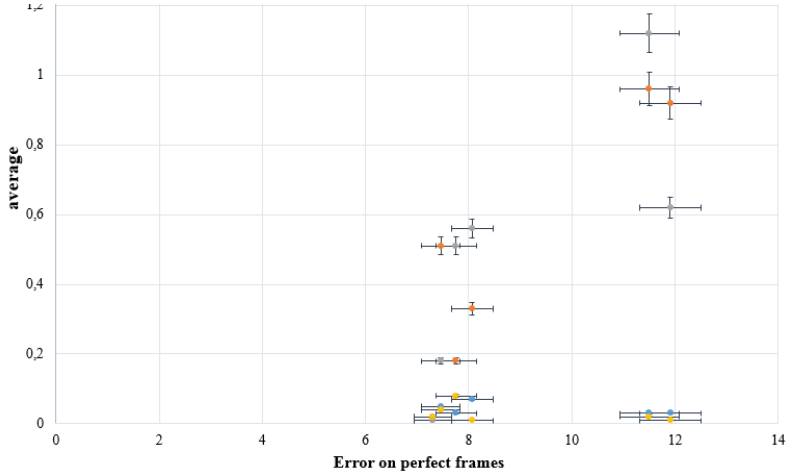


Fig. 8 Variation Error Average on Perfect Frame

VI. Conclusions and Future Work

To implement a Gaming as a Service (GaaS), an infrastructure that allows seamless experience for the user interaction is required. The number of frames per second that a mobile device processes according to a scene, facilitates interaction experience. However, the more complex a scene, a greater processing power is needed to process all data in real time. One solution to this problem is to take this responsibility to the mobile device, but the unstable network connectivity is the bottleneck.

To optimize the rendering process to a good rate according to the processing capability and network speed, it is necessary to perform a multi objective analysis. By validating different characteristics simultaneously, an efficient distribution of work is done. In the first instance, subdivision of surface is required using the octree algorithm, in order to divide the problem into less complex problems.

To model the entire process, a known and finite set of Hidden Markov Models, one for each node o domine, must be built. Each domine consists of a set of heuristics with unknown cardinality, each one of which could have different states, associated to the local responsability.

This research had only covered the volume rendering process, regardless of color, texture and lighting of the scene. However, these factors can be as crucial as the volume of the object in the interaction with the user.

In future researches, it becomes necessary to involve the algorithms associated with the shadows, textures and lighting to optimize a platform of Gaming as a Service.

References

[1] Bolívar Holman, González Rubén, Pascual Jordan, Sanjuan Oscar (2010) Assessment of learning in environments interactive through fuzzy cognitive maps, Soft Computing, vol 19, pp 1037 C1050

[2] Barab S, Thomas M, Dodge T, Carteaux R, Tuzun H (2005) Making learning fun: quest Atlantis, a game without guns, Educational Technology Research and Development, vol 53, no. 1, pp 86 C107

[3] Siu-Lung J, Chi-Wai R, Fang Y (2012) The effects of peer intrinsic and extrinsic motivation on MMOG game-based collaborative learning, Information & Management, vol 49, no. 1, pp 1 C9

[4] Gonzalez-Crespo R, Rios-Aguilar R, Ferro-Escobar, Torres N (2012) Dynamic, ecological, accessible and 3D Virtual Worlds-based Libraries using OpenSim and Sloodle along with mobile location and NFC for checking in, International Journal of Interactive Multimedia and Artificial Intelligence, vol 1, pp 62 C69

[5] Neri R, Lopez M, Bolivar-Baron H, Gonzalez-Crespo R (2023) Annotation and Visualization in Android: An Application for Education and Real Time Information, International Journal of Interactive Multimedia and Artificial Intelligence, vol 2, pp 7 C12

[6] Hoberman P, Krum D. M, Suma E. A, Bolas M (2012) Immersive training games for Smartphone-based head mounted displays, Proceedings of Virtual Reality Workshop, pp 151 C152

[7] Cai W, Leung V, Hu L (2013) A Cloudlet-Assisted Multiplayer Cloud Gaming System, Mobile Networks and Applications, vol 19, pp 144 C152

[8] Feisst M, Christ A (2004) Dynamically Optimized 3D (Virtual Reality) Data Transmission for Mobile Devices, In: Proceedings of the 2nd International Symposium on, pp 270 C274

[9] Nguyen Thu D, Peery C, Zahorian J (2001) DDDDRRaW: A Prototype Toolkit for Distributed Real-Time Rendering on Commodity Clusters, Proceedings of 15th International Parallel and Distributed Processing Symposium

[10] Bolívar H, Martínez M, Gonzalez R, Sanjuan. O (2012) A multi-agent matchmaker based on hidden markov model for decentralized grid scheduler, Proceedings of 4th International Conference on Intelligent Networking and Collaborative Systems (INCoS), pp 628 C636

[11] Stone J, Saam J, Hardy D, Vandivort K, Hwu W, Schulten K, (2009) High performance computation and interactive display of molecular ortitals on GPUs and multi-core CPUs, Proceedings of 2nd Workshop on General Purpose Processing on Graphics Processing Units, pp 9 C18

[12] Anderson D, Estrada T, Taufer M, Reed K (2009) EmBIOINC: An emulator for performance analysis of BOINC projects, Proceedings of the 2009 IEEE International Symposium on Parallel & Distributed Processing, pp 1 C8

[13] Yamagiwa S, Wada K (2009) Performance study of interference on GPU and CPU resources with multiple apllications, Proceedings of the 2009 IEEE International Symposium on Parallel & Distributed Processing, pp 1 C8

[14] Thomas D, Howes L, Luk W. (2009) A comparison of CPUs, GPUs, FPGAs, and massively parallel processor arrays for random numer generation, Proceedings of the ACM/SIGDA International symposium on Field programmeable gate arrays, pp 63 C72

[15] Bethel W, Shalf J (2003) Grid-Distributed Visualizations Using Connectionless Protocols, Computer Graphics and Applications, pp 51 C59

[16] Corcoran A, Redmond N, Dingliana J (2010) Perceptual enhancement of two-level volume rendering, Computers & Graphics, vol 34, no. 4, pp 388 C397

[17] Wald I, Ize T, Parker S (2008) Fast, Parallel, and Asynchronous Construction of BVHs for Ray Tracing Animated Scenes, Computers & Graphics, vol 32, pp 3 C13

[18] Madhavan, K.P.C. Arns L.L, Bertoline G.R (2005) A distributed rendering environment for teaching animation and scientific visualization, Computer Graphics and Applications, vol 25, no. 5, pp 32 C38

[19] Taura K (2004) GXP: An Interactive Shell for the Grid Environment, In International Workshop on Innovative Architecture for Future Generation High-Performance Processors and Systems, pp 59 C67

[20] Kamoshida Y, Taura K (2008) Scalable Data Gathering for Real-time Monitoring Systems on Distributed Computing, CCGRID '08. 8th IEEE International Symposium on Cluster Computing and the Grid, pp 425 C432

[21] Mahovsky J, Benedicenti L (2003) An Architecture for Java-Based Real-Time Distributed Visualization, Visualization and Computer Graphics, vol 9, no. 4, pp 570 C 579

[22] Carmona R, Froehlich B (2011) Error-controlled real-time cut updates for multi-resolution volume rendering, Computers & Graphics, vol 35, no. 4, pp 931 C944

[23] Scheiblauer C, Wimmer M (2011) Out-of-Core Selection and Editing of Huge Point Clouds, Computers & Graphics, vol 35, no. 2, pp 342 C351

[24] Klein J, Friman O, Hadwiger M, Preim B, Ritter F, Vilanova A, Zachmann G, Bartz D (2009) Visual computing for medical diagnosis and treatment, Computers & Graphics, vol 33, no.4, pp 554 C565

[25] Shirley P, Sung K, Brunvand E, Davis A, Parker S, Boulos S (2008) Fast ray tracing and the potential effects on graphics and gaming courses, Computers & Graphics vol 32, no. 2, pp 260 C267

[26] Santos A, Teixeira J. M., Farias T, Teichrieb V, Kelner J (2012) Understanding the Efficiency of kD-tree Ray-Traversal Techniques over a GPGPU, Architecture International Journal of Parallel Programming, vol 40, no.3, pp 331 C352

[27] Steinberger M, Kainz B, Hauswiesner S, Khlebnikov R, Kalkofen D, Schmalstieg D (2012) Ray prioritization using stylization and visual saliency, Computers & Graphics, vol 36, no. 6, pp 673 C684

[28] Parker S, G., Bigler J, Dietrich A, Friedrich H, Hoberock J, et al. (2010) Optix: a general purpose ray tracing engine. ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2010, vol 29, no. 66

[29] Castanie L, Mion C, Cavin X, Levy B (2006) Distributed Shared Memory for Roaming Large Volumes, IEEE Transactions on Visualization and Computer Graphics, vol 12, no. 5 pp 1299 C1306

[30] Xing X, Jia R, Lv X (2011) A Surface Reconstruction Algorithm based on Octree and Optimized Local Smooth Degree, Procedia Engineering, vol 15, pp 3728 C3732

[31] Wald I, Havran V (2006) On building fast kd-trees for ray tracing, and on doing that in O (N log N), In: Proceedings of the 2006 IEEE Symposium on Interactive Ray Tracing, pp 61 C69

[32] Baum L, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, The Annals of Mathematical Statistics, vol 26, pp 164 C171

[33] Xhafa F, Abraham A Computational models and heuristic methods for Grid scheduling problem, Future Generation Computer System, vol 26, pp 608 C621

Holman Bolívar, PhD, is the director of Software Engineering Program at the Universidad Católica de Colombia. Professor of computer programming and Artificial Intelligence. He has participated in numerous projects I + D + I such as Senacyt and Colciencias among others. His research and scientific production focuses on serious gaming, web engineering, mobile technologies and cognitive training. He has published more than 20 works in indexed research journals and conferences.

John Velandia, holds a MSc**.** from Stuttgart University (Germany). He has been leader of IT projects in the industry, in sectors such as Education, Food an Automotive. He manages his projects based on PMI, TOGAF and ITIL. Moreover, He teaches software architecture and guide bachelor thesis at the Catholic University. He is responsible of a research junior group named GINOSKO. The fields of research area are Web technologies, Business Intelligence and IT frameworks.

Jenny Torres Engineer telematics and magister in software engineering. professor of Software Engineering Program at the Universidad Católica de Colombia. She has directed numerous projects Social Responsibility. She is a enthusiastic and leader person who has gained knowledge and experience in areas such as software engineering and project management

Elena Giménez B.S. Degree in Mathematics at Faculty of Mathematics, Universidad Complutense and PhD at Engineering School, Univerisdad de Jaén (Spain). She has published several scientific papers in international peer-reviewed journals. Nowadays, she is working at Universidad Internacional de la Rioja, as an assistant professor, proyect manager and continues doing research work in engineering and technology.

SRec and VAST: Visualizing Software with a Student-Centered Aim

Francisco Javier Almeida-Martínez and Antonio Pérez-Carrasco

Universidad Internacional de La Rioja, La Rioja, Spain

***Abstract* — This paper reviews software visualization focused on the educational environment. Software visualization is a very wide study field, so we have focused on two areas: recursion visualization and parsers' visualization. The paper contains a retrospective about what has been made on it, what lacks we have found and the solution provided by the authors: SRec and VAST, two software tools trying to make a significant difference between them and the software made before.**

***Keywords* — Software visualization, parsers' construction, recursion, educational visualizations.**

I. Introduction

SOFTWARE visualizations tools are used in different context in order to improve and make easier the students learning process. Many different kinds of tools can be used to improve student learning process and motivation. For example, social networks have a large impact in student motivation and communication [45]. However, the results obtained are not always positive, so there exist different variables, which can reduce the educational impact. Nowadays, introducing multimedia techniques can improve the learning process [48].

In this work we present two different software visualization tools to display different aspects as can be recursion and parser analyzed process. However, the development and evaluation methodology used in both cases is similar: student centered. This means that all functionalities have been included, improved or removed according the evaluations results.

The rest of the paper is structured as follow. In section II we describe the most important related work in software visualization for recursion and parser generation. In section 3 we describe SRec system. VAST is described in section 4. In section 5 the evaluation process for SRec and VAST is described. Finally in section 6 we set the conclusions and future works.

II. Related works

This section contains an introduction to software visualization and a review about how recursion and parsers are visualized by several already-made software tools, finding out lacks and shortages, and proposing a new software tool for those cases, SRec and VAST.

A. Software visualization

Software visualization is a technical tool for representing in an electronic, animated and interactive way. Most part of these representations tries to make easier the software comprehension.

The educational environment focused on computer science is one of the contexts where more software representations are used, but currently is not massive. Teachers are reluctant to adopt software and new ways to teach; they feel losing the control of the class when they use new software. Lack of evidences about visualization effectiveness is an important factor to explain why software visualization is not used in most classes.

In order to fix it, software for visualizations is created, taking usability recommendations and exhaustive analyses about what both teachers and students need.

B. Recursion visualization

Recursion visualization comprises the process of representing graphically the recursion, providing animation and interaction features. Recursion is a process or software function that requires its own service once or several times to find a solution. Every time the function is called by itself, the size of the problem is smaller, letting it to reach the base case, when the problem can be solved in an easy and direct way.

Recursion is a hard concept to be learned, help students to learn it through recursion visualization has been the main goal of a lot of software. These software applications usually use animations for describing step by step how recursion achieves to solve a problem. Student interaction is very important to make easier learning tasks [22] like algorithm analysis or debugging.

Recursion can be taught using different conceptual models. A conceptual model provides a singular representation for a concept, system or event, and must be complete, coherent and precise. For recursion, there are some conceptual models widely accepted and used for teaching recursion [19][42]. The most abstract one is the inductive model, defined as a mathematical formula where the base case is directly identified. Metaphors are very used because they make easy the identification of concepts with daily life (Russian dolls [14] or mirrors [41]).

Going deeper in computer science education, there are several conceptual models used at the classrooms. Trace is one

DOI: 10.9781/ijimai.2015.338

of them; every recursive call and its results are textually represented in a properly indented way. For multiple recursions the tree model is recommended, since it shows in a very clear way the nodes dependence. The animation runs over the tree in deep mode. We can find two kinds of trees: recursion tree (every node contains input parameters) or activation tree (every node contains input parameters and output result).

The third model shows a control stack where student can see a node for every unfinished call and the chronological order of callings. Copy-model broadens the control stack model, adding the source code into the visualization and/or the local variables in different windows or panels, properly stacked.

C. Software for recursion visualization

Within the functional paradigm, we can find applications like Kiel [11], RainbowScheme [21] and WinHIPE [29]. Kiel shows the execution of first-order logic programs through a syntax-structure tree, providing several functionalities for controlling the execution.

RainbowScheme allows students to see semantic content representations of programs coded using Scheme language. Students can see recursion tree and the stack status. The code is colored in order to identify the parts of the code shown in those visual representations. The program can be executed step by step.

WinHIPE is an IDE where user can see an expression evaluation as a process of rewriting. Expressions are displayed in a visual format through lists and trees. WinHIPE provides a set of configuration options for making a more understandable format. The steps sequence can be played entirely or partially, in automatic or manual ways.

Within the imperative paradigm, there are applications like ETV [34] and Jeliot [10], oriented to computer science, and other software like EROSI [18], Function Visualizer [15], Recursion Animator [40] and SimRecur [43], focused on recursion teaching.

ETV shows a copy of the code by every recursive sentence is executed. In every copy the current line is marked, that helps to follow the execution sentence by sentence. Jeliot provides a recursion tree and a copy-model view, similar to ETV, for Java-coded programs. It allows students to see which recursive calls are unfinished.

EROSI uses the copy model to show recursion. User can see the passive flow and the active flow, how data are transmitted and the output of the calls. The programs list is prefixed.

Function Visualizer works with Java-coded programs, showing step by step how the program is executed and opening a new window for every function call, so it is very easy for students to know which calls have not finished yet.

Recursion Animator requires recursive Pascal code to work. It uses the copy model, opening a new window for every called function. User can navigate forwards and backwards for repeating some parts of the execution.

SimRecur window contains several views like recursion tree, copy-model representation, stack and information about input parameters.

TABLE I
CONCEPTUAL MODELS USED TO DISPLAY RECURSION

	Copy (var.)	Copy (code)	Trace	Recurs. tree	Activ. tree	Control stack	Colored code
EROSI	X						
ETV		X	X	X			
Function Visualizer	X	X					
Jeliot 3	X	X		X			
KIEL				X			
RainbowScheme				X		X	X
Recursion Animator	X						
SimRecur		X		X		X	

Most used conceptual models are copy-model and recursion tree as is shown in Table . Both of them are capable of showing how recursive programs are executed step by step and show the whole story of the execution. However, most programs chose only one or two models, so they offer a limited vision of recursion.

The previous programs work with animations, it is essential to make a representation for recursion. User can see step by step or in an automatic way how data are transmitted or calculated. However, just a few programs allow users to go backwards in order to repeat parts of the visualization.

Besides, interaction possibilities were not widely exploited by them. Sometimes ask for more information, move the animation to a determined point, or mark nodes may be interesting actions impossible to do with this programs. As we said, interaction makes easier learning tasks [22], so this is a gap that must be filled in order to improve how recursion can be shown and taught in educative environments.

D. Parsers visualization

Parser visualization is another example of software visualization. There exists some new methodologies as for example ART [33], TML [25] and HAS [39] used to teach language processors/compilers courses. Besides, these alternatives consider the use of visualizations/animations tools with the aim of improving students learning [20]. These visualizations tools can be classified in two different groups. On one hand we have those with a theatrical aim, so their functionalities and characteristics indicate that they can be only used in a studying environment. On the other hand, we have other tools with a practical aim, so the visualizations/animations generated are oriented to improve parser development. One of the most representative tools in the theatrical group could be JFLAP [32] because it allows visualize/animate FDAs used in lexical analysis within compilation process. Other tools in this group could be THOTH [17] or BURGRAM [16]. In the second group we have those tools, which main characteristics are that they allow visualizing parsers generation and generating analyzers for a specific language. However, within this group we have distinguished three subgroups of tools, so although they have a practical motivation, visualizations generated are oriented for

different types of users. The first subgroup contains those tools which do not generate parsers but have a strict relationship with a specific language being able to display its behavior. Some examples could be: ICOMP [24], VisiClang [30], APA [33] and Tree Viewer [35]. In the second subgroup it could be those tools, which can generate parsers but their visualizations/animations are oriented for advance users. One example of these tools could be VCOCO [31]. Finally, in the third subgroup it could be those tools, which are able to generate both animations and parsers but any user can use their visualizations. Some examples of these tools could be LISA [27], VisualYacc [F128], ANTLRworks[12], Jaccie [23] and GYacc [25].

In Table is shown a summary of these tools analyzing their main characteristics.

TABLE II
SUMMARY OF SYNTAX ANALYSIS VISUALIZATIONS TOOLS. COLUMN PARSER TYPE INDICATES THE PARSER'S OUTPUT TYPE. COLUMNS ALGORITHM AND TREE INDICATE IF THESE STRUCTURES ARE DISPLAYED. AVAILABILITY COLUMN INDICATES IF THE TOOL CAN DOWNLOAD.

	Parser type	Algorithm	Tree	Generation	Availability
ICOMP	LL(1)	X	X		
VisiClang	LL(1)	Grammar			
APA	SLR/ LL(1)	X	X		
Tree Viewer			X		
VCOCO	LL(1)	Grammar		CocoR	
AnaGram	LALR(1)	Grammar		Own	
CUPV	LALR(1)	X		Cup	
LISA	SLR(1)/ LALR(1)/ LL(1)		X	Own	X
Visual Yacc	LR		X	Yacc	
ANTRLWorks	LL(k)		X	ANTLR	X
JACCIE	LL(1)/ SLR(1)/ LALR(1)		X	Own	X
GYacc	LALR(1)		X	Yacc	

Analyzing Table we can see that all those tools, which can generate a parser, depends on an own notation or sometimes use a specific generation tool. This way of work makes difficult to use any of these tools with an educational aim because the visualizations are dependent of a generation tool.

As there do not exist any system, which implement all fundamental characteristics (algorithm, syntax tree, etc) for parser visualization, if we want to display a new dimension, it would be necessary to change both generation and visualization tools. From the students point of view it means to learn how to use two different tools with different characteristics: syntax, parser's build process, understanding output messages (for example LR shift-reduce conflict). From the teacher point of view it means to get familiar with the new environment. As we can guess, this way of working may prevent teacher to use these tools in their lessons [28].

There exist other two aspects to take into account according to the described tools: educational evaluation and availability. Referring to the educational use of these tools, there do not exist empirical data, which analyze the impact on the students learning process. Another aspect is the availability, so it is often difficult to get one on these tools.

III. SREC. ¿VISUALIZING RECURSION?

SRec [36][37][38] is a software application developed in order to provide animated and interactive visualizations about recursive Java-coded programs. SRec is aimed to help both students and teachers in algorithms courses. This software provides the visualizations generated in an automatic way, it means that users only provide the class and the method they want to run and the input parameters they want to use. SRec compiles the class, runs the program, saves the needed internal data and creates the visual representation of the views in a few seconds. After that, users can go step by step through the recursive execution (SRec does not advance sentence by sentence but opening or closing recursive calls at a time). Users can go forwards or backwards in an automatic or manual mode. Users can go step by step or skip all the steps until closing the current active node (the active node is the last opened recursive call so far).

SRec offers through its window, shown in Fig. 4, four different views at the same time. One of them (at the left) shows the source code of the Java class loaded by the users, where they can modify it, save it and recompile it again in order to add, delete, change or correct one or several sentences. The second view (located under the code view) is the compiler view, where users can see if the changes they wrote are right or generate some compilations errors.

When users create a visualization, two views are opened. These views show two different representations of the program at the same time and these representations can be changed through the tabs they contain. Every tab let see a different representation or conceptual model.

SRec offers three conceptual models for recursive programs. Users can access to the recursion tree through the "Tree" tab, where input parameters are shown in nodes, and activation tree (see Fig. 5), where every node contains input parameters and result values.

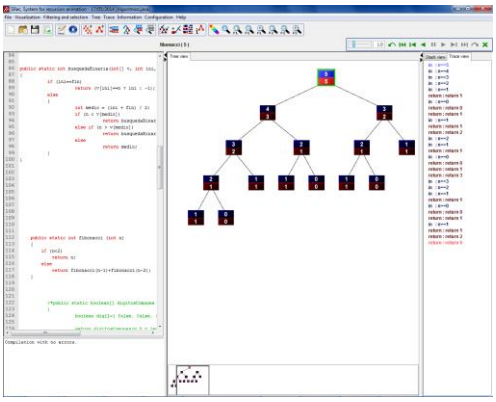


Fig. 4. Main window of SRec

This tree view contains a thumbnail representation of the whole tree, very useful for handling very big trees. The stack

view shows the unfinished calls and the dependence between them, it just contains a subset of the tree view nodes. The trace view contains a text line for every step the visualization has advanced so far. They are indented according to the depth level and the color is different for openings and closings operations.

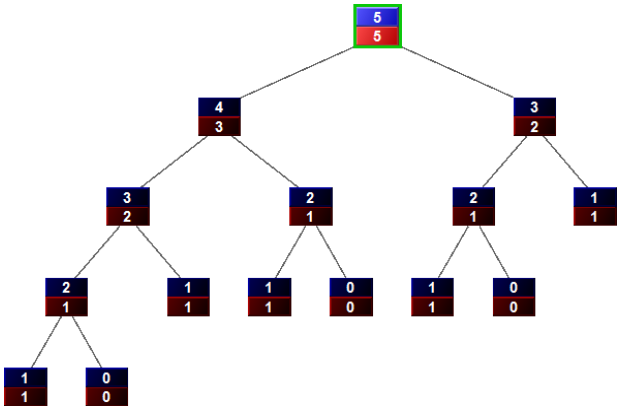


Fig. 5. Activation tree generated by SRec for Fibonacci algorithm (input: 5)

SRec provides two additional conceptual models for divide-and-conquer algorithms. These algorithms usually act on a data structure, dividing it in order to make a directly solvable problem. SRec supports arrays and matrixes, so it can represent them in an "extended" tree, where a small representation of the data structure is added to every node. Besides, two additional tabs are activated for showing a chronological view and a structure view.

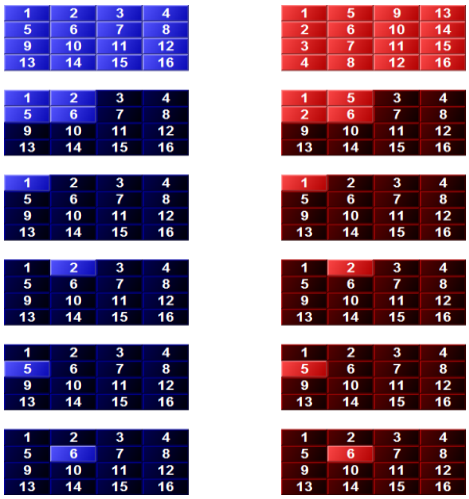


Fig. 6. Chronological view (not complete) generated by SRec for divide-and-conquer algorithm (transposing a matrix).

The chronological view (see Fig. 6) shows all the statuses of the structure along the execution in a chronological order. At the left, input values are shown; at the right, returned values are displayed. User can see how the algorithm is working step by step on the different zones of the structure. The structure view (see Figure 7) always shows the current status of the structure, adding lines for arrays or boxes for matrixes below the structure for marking the affected areas by every unfinished call of the program. These views darken the areas affected by finished recursive calls and the areas not affected by any already open call.

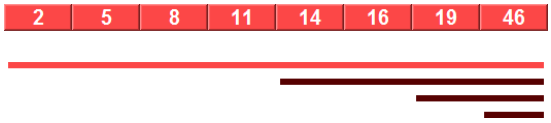


Figure 7: Structure view generated by SRec for binary search (searching for number 46)

The format of visualization is configurable. Colors for input parameters and output values in nodes, how source code is colored, location and separation of nodes in views, edges shape… can be configured by the user. This feature is important in order to adequate the visualization to different environments (monitor, projector, big or small screens, strong or weak illumination…).

SRec allows save on disk the current visualization. SRec can open it in another work session in a very fast way, just choosing the saved file. The visualization will be restored at the same point, with the same data, and the same format. This is a very useful functionality for teachers; they can load in a few seconds a lot of previously saved examples created by themselves.

IV. General Visualization Model and VAST

Once we have analyzed the general limitations of the parser's visualization generation tools, we plan to create a system with the following objectives:

1) **Independence from the parser generation tool.** Building an independent visualization tool would make easier to use it in educational context.
2) **Display all fundamental structures.** It should be possible to display fundamental structures. New views should be added without effort.
3) **Review of the educational impact.** The generated animations should improve or make easier the students learning process.
4) **Availability.** This tool should be easy to download.
5) **Syntax error recovery.** Generated animations should display how the parser recovers from a syntax error.
6) Building generic syntax analysis visualizations.

Once analyzed the visualizations built by the generation tools of the compilation process, we present the design of a generic model to visualize/animate the compilation process. The main objective of this model is to set the base to develop a generic tool in order to solve the limitations found. The generic model can be divided in different independent submodels with certain functionalities.

In Fig. 8 we show a general scheme of this model. As we can see there exist four submodels: submodel of language processing, submodel of visualization, submodel of animation and submodel of interaction. The generalization process, usually needs a module to interpret the intermeddle actions. For example a generalization process for cartography visualization needs a module to interpret the information [46].

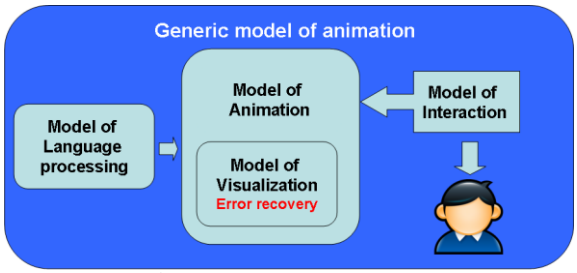


Fig. 8. Generic model of animation

Submodel of language processing. The objective of this submodel is to achieve the independence between the generation tools and the visualization tools. In this situation it is necessary to analyze the actions performed by the parsers during the execution in order to generate an intermeddle representation. Due to this intermeddle representation; it is necessary to insert extra information in the user's specifications. As result, this model should perform two different tasks:

1) Annotation process: modifications to get information during the parser's execution. This is usually performed in educational visualizations. The software platform does not determine if the annotation process is necessary or not. For example, to use the augmented reality it is needed to include new information in the real time image [47].

2) Generate a intermeddle representation.

As we have said in previous sections, as there exist different generation tools and different types of parsers, it is necessary to consider two different levels of independence. The first level has to be with working with different generation tools. The second level means that the model should be able to interpret the actions performed by the parsers.

Submodel of visualization. Using an annotated parser, its execution allows to obtain an intermeddle representation containing the visual representation. The task of this submodel is to interpret this representation and build its visual representation. In the compilation process we can highlight some internals structures as the syntax tree, parser stack, etc. Due to this, the syntax tree is the main visualization built by this submodel. Besides, other structures as the input stream, the stack, grammar and the actions performed are displayed. All these views should be synchronized. One important characteristic of this model is the expandability, so it should be easy to add the visualization of a new structure.

Submodel of animation. Its main task is to animate the static visualizations generated by the submodel of visualization. As one of the most important structures is the syntax tree, the main task of this submodel is to animate its building process keeping the synchronization between all views. The animation process should distinguish between LR and LL parser because the syntax tree and the stack have different behaviors. For a LR parser when it detects a terminal node it is added into the stack. If it detects a non terminal node, it performs a reduction deleting nodes from the stack. When working with a LL parser, if a terminal node is detected, it is removed from the stack. For a non terminal node, a derivation is performed, which means that the rule's consequent is inserted into the stack.

Submodel of interaction. This submodel is responsible of the interaction with the user. In order to this, we can distinguish two functionalities: make easier the animations creation and allow working with the generated visualizations.

VAST. Visualizer of the Syntax Tree

VAST is the result of implementing the generic model of visualization and each submodel. In Fig. 9 we can see VAST's main window. The syntax tree is the main visualization. On the right we have the input stream and the parser grammar. In the bottom we have the parser's stack; the log for action performed (only used during the animation process) and a global view to navigate throw the syntax tree.

V. USABILITY-EDUCATIONAL EVALUATIONS

SRec and VAST have been subjected to several evaluations about usability and educational effectiveness. Next we summarize the processes and the obtained results.

A. Usability evaluation process for SRec

SRec has been evaluated five times about usability. The main principles used to develop SRec were: easy installation, easy learning to use it, and efficiency approach when user is working.

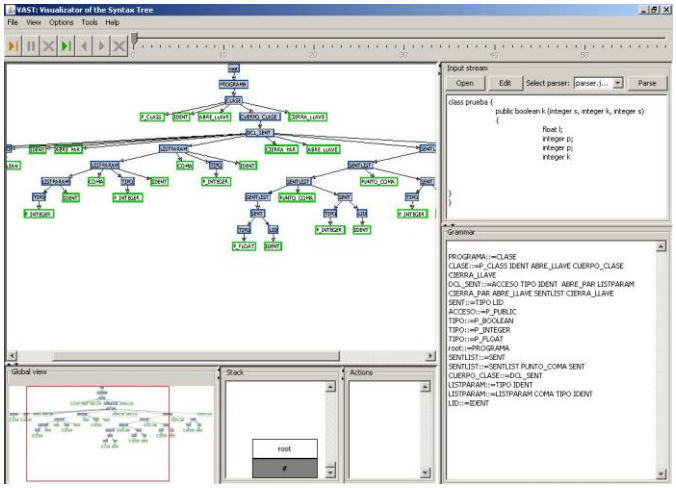


Fig. 9. VAST main user interface.

The basic schema of each evaluation session was:

- Teacher demonstration: the teacher shows to students how to create visualization from a source code for a few minutes.
- Familiarization tasks: students do some light tasks in order to get a first contact with SRec.
- Didactical exercise: one exercise must be done for the course using SRec. These exercises asked for analyzes, debugging, design or creation of an algorithm.
- Questionnaire: a questionnaire was provided to students

in order they to fill it with their opinion about SRec.

The fourth session was made using two days (taking data from both of them) and the last session took three classes (we took data only from the last day). **¡Error! No se encuentra el origen de la referencia.** contains the marks for several general questions about SRec (the minimum value was 1 and the maximum value was 5).

TABLE III
SREC SCORES FROM USABILITY EVALUATIONS

	Session 1	Session 2	Session 3	Session 4	Session 5
SRec is easy to use	3.88	4.50	4.20	4.19	3.94
General quality of SRec to analyze Recursion	3.38	4.29	4.00	-	3.84
I like SRec	3.63	4.26	3.95	3.98	3.84
Number of opinions	7	28	21	28/19	49

According to the **¡Error! No se encuentra el origen de la referencia.**, marks were growing up while the same functionality was improved (working in a more stable way and with a better interface), so marks were growing up from session 1 to session 2.

After that, new functionalities and possibilities were added, making harder how to learn to use SRec, and the asked tasks were more difficult too, so marks were lightly going down.

Table **¡Error! No se encuentra el origen de la referencia.** contains the marks (fifth evaluation) for several important SRec features.

TABLE IV
OTHER SREC SCORES FROM USABILITY SESSIONS

Animation controls	3.96
Activation tree view	4.00
Visualization generation process	4.20
Chronological view	3.86
Structure view	3.63
Menu structure	4.00
Window elements interaction (panels, scrolling…)	3.55

The questionnaires gave some additional data through open questions. The suggestions made by students were changing in the different evaluation sessions, due to SRec was adding some previous suggestions and adding new functionalities. Some of them partially changed the way how SRec must be used and it could provoke some new issues and the lowering of the marks.

In the fifth evaluation, 30% of students did not provide any functional suggestion for SRec and only a third part of students would delete some functionality of SRec. When the questionnaire asked for the hardest parts to be learnt, 18.4% answered saying that SRec was easy to use (positive answers), and 49% did not answered the question; just 32.7% talked about dark aspects of SRec. 20.4% of students spontaneously used SRec to study or to make activities for the course, as they say in the questionnaire.

B. VAST evaluation process

In case of VAST, once analysed different visualization tools we realized that none of them was evaluated in an educational way. As there do not exist empirical results, maybe those tools are used without adapting to students' needs. From the educational point of view, they could have a negative impact in students' learning process.

According to VAST, we divided the implementation process in different stages. Once finishing one stage we planned different evaluations of the generic model. Due to this, the evaluation process has been iterative, allowing adapting the functionalities to students needs. Two different types of evaluations have been executed: interactive-usability [3][4][5][9][6] and educational [2][7][8][3].

The interactive-usability evaluations had a double intention: evaluator observation and students opinion. The evaluator's observation evaluations allowed to study and annote how students used the tool and detect the main problems. The opinion evaluations were focus on asking students about the experience when using the tools. Educational evaluations have allowed observing the impact on student's learning process when using these tools. Due to this, we designed pre and post knowledge tests according to Bloom's taxonomy. The methodology used is similar in all evaluations.

Although we have distinguished three different types of evaluations, they have been performed at the same time. This means that one educational evaluation consists of observations, usability tests and knowledge tests. For the usability and educational impact evaluations, students have been divided in different groups (control and treatment) [13] randomly. Groups were balanced using a knowledge pretest.

The first evaluation of VAST was focus on observing how students worked with the tool and the problems they had. In the second evaluation we compared VAST and ANTLRworks, obtaining results in favour ANTLRworks according to usability. In this case we got significant statistical differences in educational impact (synthesis level in Bloom's taxonomy) in favour VAST. From these results we planned a global integration in VAST in order to improve its usability. Once finished the development we performed another usability evaluation. This process continued during all the development process.

VI. CONCLUSIONS AND FUTURE WORKS

This paper contains a review about software visualization aimed to students. They are not massively used in the classrooms and some causes were explained. Software visualization can be divided in several areas, and two of them were exposed in detail here: recursive programs and parsers.

For recursive programs, the most used conceptual models were commented and the most relevant software for recursion visualizations was briefly reviewed. The lack of strong interaction features was one of the main conclusions joined to the limited vision of recursion given by most of the existing

software. There, SRec marks a difference with its interaction features and its wide range of visualizations for recursion.

SRec is totally functional; it allows improve teachers' job, and students mostly like it. However, some ways to improve it have been discovered. The main future works are:

- Educational effectiveness evaluations for testing whether SRec really helps students to learn how recursion works.
- Interface simplification for making a more agile SRec.
- Predictive mode for SRec, allowing students to fill empty nodes with their expected values to check if they understood the algorithm they are visualizing.
- Enlarging the supported algorithm-design techniques in order to show specific views for them (dynamic programming, for example).
- Study several data taken from the fifth evaluation session about students' usage in order to figure out how students work with SRec. This may make possible create students profiles and adapt SRec to them.

About parsers' visualization, this paper reviews some new methodologies focused on processors and compilers teaching. There are several software tools, catalogued into two main categories: tools with a theatrical aim for educational context and other software oriented to parsers development.

This second group contains three categories. The first one is for software aimed only to visualize how a parser behaves, not to generate it. The second one contains software that can generate parsers and their visualizations are oriented to advanced users. The last category is for software able to generate both animations and parsers to a wide target of users.

The main drawback for these software tools is that they depend on a generation tool, so it is difficult to use them in an educational environment, so students have to use two different tools to understand the whole process and teachers may be reluctant to use these tools for their classes. Here is where VAST changes the paradigm.

The development process with VAST has not finished. We plan to finish the integration of syntax error recovery visualization. As described in bibliography, the syntax error recovery is usually known as a "black art" [1]. From the students' point of view, it is one of the most difficult parts to understand in parser construction. So, if a system can display this process, maybe students can learn it easily.

REFERENCES

[1] Aho, A.V., Lam, M.S., Sethi, R., Ullman, J.D. (2007). Compilers: Principles, Techniques, and Tools, Prentice Hall.

[2] Almeida-Martínez, F.J., Urquiza-Fuentes, J. (2009). Long term evaluation using VAST. Technical report. Universidad Rey Juan Carlos, 2009 1988-8074, 2011-13.

[3] Almeida-Martínez and Urquiza-Fuentes, J. (2011). Building LL(1) grammars and parsers using VAST and ANTLR. Technical report. Universidad Rey Juan Carlos, 2011 1988-8074, 2011-10.

[4] Almeida-Martínez, F.J., Urquiza-Fuentes, J. (2011). Preliminary evaluation of the syntax error recovery visualization using VAST. Technical report. Universidad Rey Juan Carlos, 2011 1988-8074, 2011-08.

[5] Almeida-Martínez, F.J., Urquiza-Fuentes, J. (2011). VAST. Evaluation of the execution-visualization functional integration. Technical report. Universidad Rey Juan Carlos, 2011 1988-8074, 2011-07.

[6] Almeida-Martínez, F.J., Urquiza-Fuentes, J. (2009). Teaching LL(1) parsers with VAST- An Usability Evaluation. Technical report. Universidad Rey Juan Carlos, 2009 1988-8074, 2009-01.

[7] Almeida-Martínez, F.J., Urquiza-Fuentes, J., Pérez-Carrasco, A. (2011). Syntax error recovery visualization. Evaluation using VAST and Cup. Technical report. Universidad Rey Juan Carlos, 2011 1988-8074, 2011-12.

[8] Almeida-Martínez, F.J., Urquiza-Fuentes, J., Pérez-Carrasco, A. (2011). Building LR(1) grammars and parsers using VAST and Cup. Technical report. Universidad Rey Juan Carlos, 2011 1988-8074, 2011-11.

[9] Almeida-Martínez, F.J., Urquiza-Fuentes, J., Pérez-Carrasco, A. (2011). Building LL(1) parsers with VAST. Second usability and educational evaluation. Technical report. Universidad Rey Juan Carlos, 2011 1988-8074, 2011-06.

[10] Ben-Bassat, R., Ben-Ari, M., Uronen, P.A. (2003). "The Jeliot 2000 program animation system". Computers & Education, 40(1), pp.1-15.

[11] Berghammer, R., Milanese, U. (2001). "KIEL: A computer system for visualizing the execution of functional programs". Proceedings of the International. Workshop on Functional and (Constraint) Logic Programming 2001, Technical. Report No. 2017, pp. 365-368, Kiel, Germany: University of Kiel.

[12] Bovet, J., Parr, T. (2008). ANTLRWorks: an antlr grammar development environment. Software: Practice and Experience, 38(12):1305–1332.

[13] Cohen, L., Manion, L., Morrison, K. (2000). Research Methods in Education. RoutledgeFalmer, 2000.

[14] Dale, N.B., Weems, C. (1991): "Pascal", 3rd ed. Lexington, MA: D.C. Heath.

[15] Dershem, H.L., Erin Parker, D., & Weinhold, R. (1999): "A Java function visualizer". Journal of Computing in Small Colleges, Vol. 15.

[16] García Osorio, C., Gómez-Palacios, C., García-Pedrajas, N. (2008). A Tool for Teaching LL and LR Parsing Algorithms. In ITiCSE '08: Proceedings of the 13th Annual Conference on Innovation and Technology in Computer Science Education, pages 317–317, New York, NY, USA. ACM.

[17] García Osorio, C., Mediavilla Sáiz, I., Jimeno Visitación, J., García Pedrajas, N. (2008). Teaching Push-Down Automata and Turing Machines. In ITiCSE '08: Proceedings of the 13th Annual Conference on Innovation and Technology in Computer Science Education, pages 316–316, New York, NY, USA. ACM.

[18] George, C.E. (2000). "EROSI: Visualising recursion and discovering new errors". Proceedings of the 31st SIGCSE Technical Symposium on Computer Science Education, SIGCSE'00, pp. 305-309, New York, ACM Press

[19] Haynes, S.M. (1995): "Explaining recursion to the unsophisticated". ACM SIGCSE Bulletin, Vol. 27(3), pp. 3-6 y 14.

[20] Hundhausen, C.D., Douglas, S.A., Stasko, J.T. (2002) A Meta-Study of Algorithm Visualization Effectiveness. Journal of Visualization Langauge and Computer animation, 13(3):259–290.

[21] Jehng, J.-C.J., Tung, S.-H.S. & Chang, C.-T. (1999): "A visualization approach to learning the concept of recursion". Journal of Computer Assisted Learning, Vol. 15, pp. 279-290.

[22] Kehoe, C., Stasko, J., Taylor, A. (2001): "Rethinking the evaluation of algorithm animations as learning aids: an observational study". International Journal of Human-Computer Studies, Vol. 54(2), pp. 265–284.

[23] Krebs, N., Schmitz, L. (2004). Jaccie Handbook. UniBw München.

[24] Kristy, A., Robert, H., Wayne, Y. (1998). Design and implementation of the UW illustrated compiler. In PLDI '88: Proceedings of the ACM SIGPLAN 1988 conference on Programming Language Design and Implementation, pages 105–114, New York, NY, USA. ACM.

[25] Ledgard, H.F. (1971). Ten mini-languages: A study of topical issues in programming languages. ACM Computing Surveys, 3(3):115–146.

[26] Lovato, M.E., Kleyn, M.F. (1995). Parser visualizations for developing grammars with yacc. In SIGCSE '95: Proceedings of the twenty-sixth SIGCSE technical symposium on Computer science education, pages 345–349, New York, NY, USA. ACM.

[27] Mernik, M., Zumer, V. (2003). An educational tool for teaching compiler construction. In IEEE Transactions on Education, volume 46, pages 61–68.

[28] Naps, T.L., Rossling, G., Almstrum, V., Dann, W., Fleischer, R., Hundhausen, C., Korhonen, A., Malmi, L., McNally, M., Rodger, S., Velázquez-Iturbide, J.A. (1998). Exploring the role of visualization and engagement in computer science education. In ITiCSE-WGR '02: Working group reports from ITiCSE on Innovation and Technology in

Computer Science Education, pages 131–152, New York, NY, USA, 2002. ACM.A. ACM.
[29] Pareja, C., Urquiza, J., Velázquez, J. Á. (2007): "WinHIPE: an IDE for functional programming based on rewriting and visualization". ACM SIGPLAN Notices, Vol. 42(3), pp. 14–23.
[30] Resler, D. (1990). VisiCLANG. A Visible Compiler for CLANG. ACM SIGPLAN Notices, 25(8):120–123.
[31] Resler, D., Deaver, D. (1998). VCOCO: A Visualisation Tool for Teaching Compilers. In ITiCSE '98: Proceedings of the 6th annual Conference on the Teaching of Computing and the 3rd annual Conference on Integrating Technology into Computer Science Education, pages 199–202, New York, NY, USA. ACM.
[32] Rodger, S.H., Lim, J., Reading, S. (2007). Increasing interaction and support in the formal languages and automata theory course. In ITiCSE '07: Proceedings of the 12th annual SIGCSE conference on Innovation and technology in computer science education, pages 58–62, New York, NY, USA, 2007. ACM.
[33] Sami, K., Yanti, S. (1998) Animating parsing algorithms. In SIGCSE '98: Proceedings of the twenty-ninth SIGCSE Technical Symposium on Computer Science Education, pages 232–236, New York, NY, USA. ACM.
[34] Terada, M. (2005): "ETV: a program trace player for students". Proceedings of the 10th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, ITiCSE'05, pp. 118-122, New York, ACM Press.
[35] Vegdahl, S.R. (2000). Using visualization tools to teach compiler design. In Proceedings of the second annual CCSC on Computing in Small Colleges Northwestern conference, pages 72–83, US. Consortium for Computing Sciences in Colleges.
[36] Velázquez, J. Á., Pérez, A., Urquiza, J. (2008): "A design of automatic visualizations for divide-and-conquer algorithms". Electronic Notes in Theoretical Computer Science (2009), Vol. 224. Proceedings V Program Visualization Workshop (PVW 2008).
[37] Velázquez, J. Á., Pérez, A., Urquiza, J. (2009): "Interactive visualization of Recursion with SRec". 14th ACM-SIGCSE Annual Conference on Innovation and Technology in Computer Science (ITICSE 2009), póster.
[38] Velázquez, J. Á., Pérez, A., Urquiza, J. (2008): "SRec: An animation system of recursion for algorithm courses". 13rd ACM-SIGCSE Annual Conference on Innovation and Technology in Computer Science (ITICSE 2008), pp.225-229.
[39] White, E., Sen, R., Stewart, N. (2005) Hide and show: using real compiler code for teaching. In Proceedings of the 36th SIGCSE technical symposium on Computer science education, SIGCSE '05, pages 12–16, New York, NY, USA. ACM.
[40] Wilcocks, D., & Sanders, I. (1993): "Animating recursion as an aid to instruction". Computers & Education, Vol. 23, pp. 221-226.
[41] Wirth, N. (1976): "Algorithms + Data Structures = Programs". Englewood Cliffs, NJ: Prentice Hall (1976).
[42] Wu, C.-C., Dale, N.B., & Bethel, L.J. (1998): "Conceptual models and cognitive learning styles in teaching recursion". Proceedings of the 29th SIGCSE Technical Symposium on Computer Science Education, SIGCSE'98, pp. 292-296, New York, ACM Press.
[43] Wu, C.-C., Lee, Lin, J.M.-C., & Hsu I.Y-W. (1996): "Closed laboratories using SimList and SimRecur". Computers & Education, Vol. 28, pp. 55-64.
[44] Elizabeth L. White, Jeffrey Ruby, and Laura Denise Deddens. Software visualization of lr parsing and synthesized attribute evaluation. Software: Practice and Experience, 29(1):1–16, 1999.
[45] J.A.Cortés, and J.O.Lozano, Social Networks as Learning Environments for Higher Education, International Journal of Artificial Intelligence and Interactive Multimedia, vol. 2, issue Special Issue on Multisensor User Tracking and Analytics to Improve Education and other Application Fields, no. 7, pp. 63-69, 09/2014.
[46] Lorezo, W., R. Gonzalez-Crespo, and A. Castillo-Sanz, A Prototype for linear features generalization, International Journal of Artificial Intelligence and Interactive Multimedia, vol. 1, issue A Direct Path to Intelligent Tools, no. 3, pp. 59-65, 12/2010.
[47] Neri, R. B., G. M. Lopez, H. Bolivar-Baron, and R. Gonzalez-Crespo, Annotation and Visualization in Android: An Application for Education and Real Time Information, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 2, issue Regular Issue, no. 2, pp. 7-12, 06/2013.
[48] H. Bolivar-Baron, R. Gonzlález-Crespo, J. Pascual-Espada and O. Sanjuán-Martínez, Assessment of learning in environments interactive through fuzzy cognitive maps, Soft Computing, vol 19, issue 4, pp. 1037-1050.

Francisco Javier Almeida-Martínez was born in 1984 Las Palmas, Spain. He graduated in Computer Science in 2007 in Rey Juan Carlos University, Spain. In 2011 finished his PhD in software visualization in the Rey Juan Carlos University. Nowadays he works as a Software Engineer for payment services and at La Rioja International University as a lecturer.

Antonio Pérez-Carrasco was born in 1983 in Madrid, Spain. He graduated in Computer Science in 2008 in Rey Juan Carlos University, Spain. In 2011 finished his PhD in software visualization in the Rey Juan Carlos University. Nowadays he works as a Software Engineer for payment services since 2012 and at La Rioja International University as a lecturer since 2012.

Formalization of Event Perception and Event Appraisal Process

Shikha Jain and Krishna Asawa

Department of Computer Science, Jaypee Institute of Information Technology, Noida, India

***Abstract* — Integration of emotion in a virtual agent is a topic of research to depict human-like behavior in a simulated environment. For the last few decades, many researchers are working in the field of incorporating emotions in a virtual agent. In the emotion model, the behavior of an agent depends upon how the event is perceived by the agent with respect to the goal. Hence, perception of the event while considering the past experience, importance of event towards achieving goal, agent's own capabilities and resources is an important process which directly influences the decision making and action selection. The proposed models, till date, are either too complex to adapt or are using a very few parameters to describe the event. So, in this paper, we propose an extension of perception process in an existing emotion model, EMIA and suggest the formalization of event perception and appraisal processes to make it adaptable. This has been carried out using five parameters for event description along-with fuzzy logic which makes the process more effective yet simple.**

***Keywords* — Emotions, Emotion modeling, Event Perception, Event Appraisal.**

I. Introduction

THE world presented to human being is dynamic and continuously changing which persists momentarily and gets change. A state of world at a particular moment of time is perceived by human as a discrete event. There can be multiple angles to look at the happened event. But a person (an agent) perceives the event only from the angle which is more agreeable to his vibes. And hence, different person perceives the same event differently. Also, same person can perceive the same event differently, if occurred at two different moments of time. Moreover, perception is also influenced by a person's current emotion state. In an experiment conducted by OCC [1], they divided the participants in two groups. For the first group they played some "happy" song and for the second group, they played a "sad" song. Then, they asked all participants to guess the steepness of a hill. The participants from group two estimated it as more steeper than estimated by group one. This shows that person with negative emotions percept the environment negatively and hence concludes that emotions influence our perception as well. Nevertheless, our emotions, which are stored as past experience in our long term memory, also influences our perception of an event. The way we percept the event, changes our emotion state and hence our action. For example, suppose a 4 yrs old kid is playing with a toy and broke it. Now, he may percept the event as good or bad, depending upon his past experience in similar event. If, in the past, his mother has bought a new toy to replace this broken one, then he will percept the event as a good event in the hope that he will get a new toy and will express his happiness. On the other hand, if, in past, his mother has scolded him for breaking the toy, then he will percept the same event as bad event in the fear that he will be scolded by his mother and will express sadness. Moreover, one's capability, resources available and importance of event towards achieving the goal also play an important role in visualizing an event occurred in external environment.

On occurrence of an event in the external environment, it is sensed, perceived and is recorded in the perceptual memory which successively elicits the emotion and then decides upon an action. In the percept process, instead of recording the happened event as it is, event is perceived on inquiring the following queries:

1. How much important the event is, towards achieving the goal?
2. From past experience, what belief (good or bad) does agent has for the object involved in the event?
3. Are sufficient resources available with the agent to handle the situation?
4. How much capable the agent is to deal with the event?

In conclusion, the perception of an event depends on various factors: the sensed information, goal importance, the past experience about the object involved in the event, agent's capabilities and the resource available.

Classically, an emotional intelligent agent follows the percept-appraise-elicit-act cycle. Since, the entire cycle is depending upon the information extracted from perception process; it is a crucial process that needs to be addressed carefully. Incorrect information at perception process may lead to inconsistent action. So, in this paper, we mainly focus on the perception process. For that, we propose an extension of perception process in an existing emotion model, EMIA. Furthermore, we are suggesting the formalization of event perception and appraisal processes to make the model adaptable. As human cognitive processes are ambiguous in nature and can be defined using soft boundaries only, the use of linguistic variables and fuzzy sets allow us to model the human processes in the way, they are actually formed,

DOI: 10.9781/ijimai.2015.339

organized and handled. Moreover, the constraints with soft boundaries can be obtained by using fuzzy rules.

The paper is organized as follows. In section 2, related work is discussed in brief. Section 3 describes the emotion model, EMIA. Section 4 explains our proposed work, in detail. In section 5, scenario based illustration is demonstrated. Section 6 concludes the paper with a discussion.

II. Related Work

For the past few decades, many researchers started believing that emotions are important for every human cognitive process and the research in emotion has shown a rapid growth. Emotion research involves emotion capturing from different modalities using feature extraction [3][8][12][14-18][21], emotion modeling [2,4-5,7,10-11,19-20,22] and emotion expression.

El-Nasr et al [11] proposed a computational and adaptive model of emotions based on event appraisal called FLAME. But this model uses only two parameters: the impact of an event on a goal and the importance of this goal for the agent to decide the desirability of the event. Eva Hudlicka [4] proposed a symbolic affective cognitive architecture MAMID (Methodology for Analysis and Modeling of Individual Differences) that models a number of representative constructs and processes to implement many traits (temperament, personality, emotional style) and states (emotion, affective states, moods) based behavioral phenomena. The theory of Emotion and Adaptation of Smith and Lazarus was formalized by Gratch and Marsella [7] into EMA, a model to create agents that demonstrate and cope with (negative) affect. It describes perception as person's interpretation of their relationship with the environment and called is as causal interpretation.

Some of the models are very complex and hence non-adaptable while others use a very few parameters to describe the event. In conclusion, we are proposing a simple yet more effective perception process. For that, we are using the concept of fuzzy logic to maintain the ambiguous and flexible nature of the human processes. Furthermore, five parameters: goalImp, expectedEventResult, resource, sudden and belief (described in detail in section 4) are used to describe an event and to make all subsequent processes domain independent.

III. EMIA: Emotion Model

EMIA[20] is a fuzzy rules based computational model for emotional intelligent agent situated in a virtual environment. It is a generic model, loosely based upon three cognitive theories of emotions: OCC [1], Scherer [9] and Roseman et al. [6]. It uses five appraisal variables <*desirable, expectedness, outcomeProbability, causeHarm, and suddenness*> to evaluate an event and elicit the primary emotions <*happy, sad, anger, fear, and surprise*> (as defined by Ekman [13]. The highlighted features of the model are the event appraisal; emotion elicitation; learning; emotion transition; emotion regulation and decay process. Furthermore, the mental state of an agent is represented in terms of belief, goal and action.

It follows three-layer architecture, as shown in Figure 1. The first layer is where the agent interacts with the external environment. Here, the events occur and the behavioral action takes place. The second layer is to convert the domain-specific knowledge into domain-independent information for further processing and vice versa. The third layer is basically responsible for emotion elicitation, transition and regulation. Firstly, an event is perceived from external environment and is recorded in the perceptual memory. Then, the event is appraised using five appraisal variables: Desirability, Expectedness, OutcomeProbability, Suddenness and CauseHarm. These appraisal variables are the linguistic variables defined using fuzzy sets as below:

D ={UD, N, LowD, MediumD, HighD }
E ={UE, N, LowE, MediumE, HighE }
S ={low, medium, high}
OP ={VUP, UP, N, P, VP}
CH ={noHarm, causeHarmL, causeHarmH, willcauseHarmL, willcauseHarmH }

Similarly, the emotions (happy, sad, fear, anger, surprise) are treated a linguistic variables defined using fuzzy set as

Happy = { VHighH, HighH, MediumH, LowH, Neutral }
Sad = { VHighS, HighS, MediumS, LowS, Neutral }
Anger = { VHighA, HighA, MediumA, LowA, Neutral }
Fear = { VHighF, HighF, MediumF, LowF, Neutral }
Surprise(-)={VHighSr(-), HighSr-, MediumSr-, LowSr-, Neutral }
Surprise(+) = { VHighSr(+), HighSr+, MediumSr+, LowSr+, Neutral }

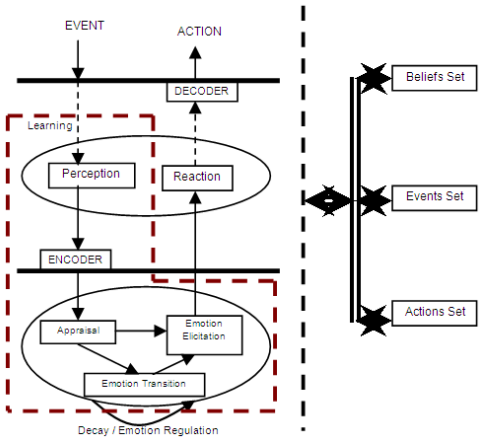


Fig. 1. Architecture of EMIA [20]

Once the event is appraised and mapped over the appraisal variables, then, the emotions are elicited based using the following rules:

if Desirability(EV_ID,TS) is A1
and

Expectedness(EV_ID,TS) is A2
and
Suddeness(EV_ID,TS) is A3
and
OutcomeProbability(EV_ID,TS) is A4
and
CauseHarm(EV_ID,TS) is A5
then
Curr_Emo1 is E1 and …. andCurr_EmoN is EN

Here A1..A5 takes the values from the respective fuzzy sets. Curr_Emo1..N is the list of multiple emotions generated simultaneously and E1 to EN are their intensities. Once the emotion is elicited, it is checked for emotion regulation. If there is next event in sequence to occur, then next emotion is elicited while considering the current emotion (emotion transition) and if no event occurs, and then the current emotion starts decaying. This cycle keeps on repeating throughout the life time of the agent.

IV. Event Perception and Event Appraisal Process and Its Formalization

Whenever an event occurs, it is sensed by an intended agent to create its own view of the external world. The agent evaluates the event in terms of various parameters, for example, liking/disliking, expectedness/unexpectedness, resources available, its own capability and the goal to be achieved. Depending upon how the agent looks at the event, emotions are triggered with some intensity. For the same event, two agents can generate different emotions depending upon how they look at the event, how they evaluate it. Hence, the emotion elicitation and action selection depends upon how an agent percepts and appraises the event. Formally, event perception and appraisal results in emotion elicitation (as shown in equation 1):

$$E_t(Aid, Evid) = Per(Aid, Evid, t) \Lambda A(Evid, t, \varphi(\psi, \beta, \varsigma)) \quad --(1)$$

Here $E_t(Aid, Evid)$ is the emotion elicited by agent *Aid* with respect to event *Evid* at time t. Emotion is elicited as a result of perception $(Per(Aid, Evid, t))$ and appraisal $(A(Evid, t, \varphi(\psi, \beta, \varsigma)))$ of the event.

In the proposed model, an agent is situated in a close environment. It has some beliefs about objects, external environment, its own capabilities and available resources. All these information is stored in the memory as belief-set. Every agent has its own belief-set. The belief-set gets updated each time an event is completed. This way it stores or updates the past experience related to a certain object. It also maintains a tag (good or bad) with every object. This tagging helps the agent to appraise the event on the basis of the past experience about the object. The goal-set keeps the set of goals need to be achieved. The goals are stored in order of priority.

Perception means receiving the event from the external environment and recording it. Same event can be viewed differently from different angles and hence can be perceived differently by different agents. For an event, multiple views can be created by different agents depending upon their beliefs. So, whenever an event occurs, it is received by the perception module in textual form. This module creates its own view of the event while considering the agent's belief-set. It assigns the event a unique event-id Evid and tags it with a time stamp. It recognizes the object/ agent responsible for the event. Then it stores the complete information in the memory as <*Evid, EV_NAME, OBJ, TS* >and generates a message containing event id, object concern and time stamp <*Eid, OBJ, TS*>. Then, it passes the message to the appraisal module. It is represented as *Per(Aid, Evid, t)* in eq(1) which means creation and representation of agents own view of the happened event.

According to the appraisal theories, the event is appraised while considering the importance of the event towards achieving the goal and agents own beliefs and available resources and then, accordingly, emotions are elicited. So, the appraisal process can be formalized as $A(Evid, t, \varphi(\psi, \beta, \varsigma))$ in eq(1) which means appraisal of the perceived event Evid while considering its importance towards the goal (ψ) and agents self-appraisal (β). Suddenness is represented by ς which measures how suddenly an event has occurred.

The symbol ψ represents the goal importance and expected outcome of the event that is whether the happened event leading us towards the goal or not. If the event contributes in achieving the goal then the event is desirable and it will generate pleasant outcome. So, the two appraisal variables: Desirability and OutcomeProbability (used in EMIA) can be mapped over ψ. Desirability is associated with each event as a measure of how desirable the event is in pursuit of the goal. OutcomeProbability is the measure of consequences (pleasant/ unpleasant) of the event with respect to the goal. If the outcome of the event is not leading us towards the goal then a value N will be assigned to it.

Agent's self-appraisal parameter (β) checks the agent's own capabilities in terms of available resources and the past experience (which is now stored in belief set as agent's belief) associated with the object/agent involved, if any. So, the two appraisal variables: Expectedness and CauseHarm (used in EMIA) can be mapped over β. Expectedness is the likelihood of that event to occur according to the agent's world view or past experience. CauseHarm is whether the current event is threatening to agent presently or in future or not. For example, if some event occurs that can cause harm but we have sufficient resources to handle the event, then the event will not cause any harm to the agent.

Since the human emotion process is very complex, it is not possible to assign a crisp value to these variables. To capture the imprecise, ambiguous and subjective information, the concept of linguistic variables is used.

The appraisal variables are defined as fuzzy set of linguistic variables in EMIA. Similarly, the fuzzy sets of linguistic variables for event perception are defined as follows:

goalImp = {High, Medium, Low, No, Neutral}
expectedEventResult = {Good, Bad, Neutral}
resource = {Yes, No, Neutral}

sudden = {High, Medium, Low, Neutral}

A set of if-else rules are designed to generate the linguistic values of the appraisal variable using the data fetched from event-set and belief-set. The generalized form of the rules is defined as follow:

if goalImp(Aid, Eid, t) is X1
and
expectedEventResult(Aid, Eid, t) is X2
and
belief(Aid, Eid, t) is X3
and
resource(Aid, Eid, t) is X4
and
sudden(Aid, Eid, t) is X5
then
D is Y1 and E is Y2 and OP is Y4 and CH is Y5and S is Y

Here X1..X5 are the linguistic values as perceived by the agent from the environment and Y1…Y5 are the linguistic values of the appraisal variables from the set.

Example: Upon seeing a snake, if an agent has a past experience (belief) that snakes are dangerous and on assessing the resources it is found that he has a stick which can be used to hit snake. Then, at the time of perception, he will assign *goalImp = no, expectedEventResult = bad, belief = Snake "BAD", resource = yes, sudden = medium.* Now, a suitable if-else rule will be fired to determine the linguistic values of the appraisal variables. After applying the rule, the value of the appraisal variables is: *D = UD, E = LowE, OP = UP, CH = willcauseHarmL, S=MediumS.*

After completing the appraisal process, this module generates a message including event id, time stamp and values of appraisal variable as <*Eid, TS, OBJ, D, E, S, OP, CH* > and passed it on to emotion elicitation module as described in EMIA [20].

V. ILLUSTRATION

This system needs to work in a variety of applications to prove its advanced features; here, we have designed a prototype model for maze game. The different scenes are designed to test the proposed model. Out of those, following scene is illustrated here.

Scenario: It's a maze game. There are many obstacles in the maze and those obstacles can block the path. In the maze, suddenly some monster can appear that can kill the agent. Moreover, the agent has two powers. Each can be used only once. Power 1 can be used to remove obstacles from the path and power 2 can be used to kill the monster. Goal of the agent is to start from the source and reach the end point of the maze. The agent moving in the maze is having the beliefs as described below.

Belief-Set: {obstacle "BAD"; destroy_obstacle "GOOD"; Monster " BAD"; Kill_monster "GOOD"; Power1; }

Event-Set: it is a set of sequence of sub-events as below:

1. Keep on moving without obstacle

[GI = High, Ex = GOOD, RA = yes, S = Neutral]
2. Obstacle appears

[GI = No, Ex = BAD, RA = yes, S = medium]
3. Destroy obstacle and move ahead

[GI = High, Ex = GOOD, RA = yes, S = Neutral]
4. Monster appears

[GI = No, Ex =BAD, RA = yes, S = High]
5. Kill monster and move ahead

[GI = High, Ex =GOOD, RA = yes, S = Low]
6. Another monster appears

[GI = No, Ex =BAD, RA = no, S = High]
7. Seek alternate path
[GI = Yes, Ex =GOOD, RA = yes, S = Neutral]
8. Reach destination.
[GI = Neutral, Ex =GOOD, RA = yes, S = Neutral]

Goal-Set: {Reach destination safely}

.Now the event appraisal if-else rules are fired one after the other to appraise the event and emotions are elicited as shown in table 1.

The model is simulated in many such scenes taken from real scenario. It is observed that the model simulates with high believability. Since we have designed exhaustive set of rules, the model is performing well in any type of given situation generated in a close loop environment

VI. CONCLUSION

Perception of the event is an important process which directly influences the decision making and action selection. Moreover, perception is influenced the past experience, importance of event towards achieving goal, agent's own capabilities and resources. The proposed models, till date, are either too complex to adapt or are using a very few parameters to describe the event. So, we propose an extension of perception process in an existing emotion model, EMIA and suggest the formalization of event perception and appraisal processes to make it adaptable. The model is based upon appraisal theories of emotions and OCC elicitation conditions.

It is a simple yet more effective perception process as it uses the concept of fuzzy logic to maintain the ambiguous and flexible nature of the human processes. Furthermore, five parameters: *goalImp, expectedEventResult, resource, sudden and belief* are used to describe an event and to make all subsequent processes domain independent. Finally, the model is simulated for a scenario designed for a maze game.

TABLE 1
THE CASE SCENARIO SIMULATION

Sno	Event	Event-Perception	Belief Base	Appraisal Variables <D, E, OP, CH, Sud>	Emotion <H, S, F, A, Sr>
1	Keep on moving without obstacle	[GI = High, Ex = GOOD, RA = yes, S = Neutral]	neutral	<HighD, MediumE, VP, noHarm, Low>	<HighH, N, N, N, N>
2	Obstacle appears	[GI = No, Ex = BAD, RA = yes, S = medium]	bad	<UD, UE, UP, willcauseHarmL, Medium>	<N, LowS, LowF, N, Medium>
3	Destroy obstacle and move ahead	[GI = High, Ex = GOOD, RA = yes, S = Neutral]	good	<HighD, HighE, VP, noHarm, low>	<MediumH, N, N, N, N>
4	Monster appears	[GI = No, Ex =BAD, RA = yes, S = High]	bad	<UD, UE, UP, willcauseHarmL, High>	<N, LowS, HighF, N, HighSr>
5	Kill monster and move ahead	[GI = High, Ex =GOOD, RA = yes, S = Low]	good	<HighD, LowE, VP, noHarm, Low>	<VHighH, N, N, N, N>
6	Another monster appears	[GI = No, Ex =BAD, RA = no, S = High]	bad	<UD, MediumE, UP, willCAuseHarmH, medium>	<N, MediumS, HighF, ***HighA***, MedSr(-)>
7	Seek alternate path	[GI = High, Ex =GOOD, RA = yes, S = Neutral]	good	<HighD, HighE, VP, noHarm, low>	<MediumH, N, N, N, N>
8	Reached destination.	[GI = Neutral, Ex =GOOD, RA = yes, S = Neutral]	good	<MedD, MedE, P, noHarm, Low>	<MediumH, N,N,N,N>

REFERENCES

[1] A. Ortony, G. Clore, A. Collins, "The cognitive structure of emotions". Cambridge university press, Cambridge, MA, 1998.

[2] A Saini, N Bhatia, B Suri, S Jain, "EmoXract: Domain independent emotion mining model for unstructured data" Seventh International Conference on Contemporary Computing (IC3), 2014, pp. 94-98

[3] B. Vishwanath, et al. "Machine learning approach for text and document mining." arXiv preprint arXiv:1406.1580 , 2014.

[4] E. Hudlicka, "Modeling Effects of Behavior Moderators on Performance". In Proceedings of BRIMS-12, Phoenix, AZ 2003.

[5] H. Jiang, J. M. Vidal, M. N. Huhns, "EBDI: architecture for emotional agents". In Proc. Of the 6th International Joint Conference on Autonomous agents and multiagent systems, 2007, pp. 1-3.

[6] I. J. Roseman, P. E. Jose, M. S. Spindel, "Appraisals of emotion-eliciting events: testing a theory of discrete emotions". Journal of personality and social psychology, 59(5), 1990, pp. 899-915.

[7] J. Gratch, S. Marsella, "A domain independent frame-work for modeling emotion". Journal of Cognitive Systems Research, 5(4), 2004, pp. 269-306.

[8] K. Asawa, and P. Manchanda, "Recognition of Emotions using Energy Based Bimodal Information Fusion and Correlation", International Journal of Artificial Intelligence and Interactive Multimedia, 2(7), 2014, pp17-21.

[9] K. Scherer, "A emotion as a multi-component process: a model and some cross-cultural data", Review of personality and social psychology, 5, 1994, pp. 37-63.

[10] M. Dastani, J. J. Meyer, "Agents with Emotions". International Journal of Intelligent Systems, 25(7), 2010, pp. 636-654.

[11] M. S. El-Nasr, J. Yen, T. R. Ioerger, "FLAME - A Fuzzy Logic Adaptive Model of Emotions". Journal of Automous Agents and Multi-agent Systems, 3, 2000, pp. 219-257.

[12] N. Bellustin, et al. "Instant human face attributes recognition system." Int. J. Adv. Comput. Sci. Appl., Special Issue Artif. Intel, 3, 2011,

[13] P. Ekman, "An argument for basic emotion". Cognition and Emotion, 6, 1998, pp 169-200.

[14] P. Kumari, and A. Vaish. "Brainwave based user identification system: A pilot study in robotics environment." Robotics and Autonomous Systems, 2015, pp. 15-23.

[15] P. Kumari, and A. Vaish. "Brainwave's energy feature extraction using wavelet transform." Electrical, Electronics and Computer Science (SCEECS), 2014 IEEE Students' Conference on. IEEE, 2014.

[16] P. Kumari, and A. Vaish. "Instant Face detection and attributes recognition." International Journal of Advanced Computer Science and Applications, 2011.

[17] P. Kumari, S. Kumar, and A. Vaish. "Feature extraction using emprical mode decomposition for biometric system." Signal Propagation and Computer Technology (ICSPCT), 2014 International Conference on. IEEE, 2014.

[18] S. J. B. Castro, R. Gonzalez-Crespo, and V. H. Medina-Garcia, "Patterns of Software Development Process", International Journal of Artificial Intelligence and Interactive Multimedia, 1(4), 2011, pp. 33.

[19] S. Jain, K. Asawa, "EmET: Emotion Elicitation and Emotion Transition Model" , Second International Conference on Information Systems Design and Intelligent Applications – 2015, Advances in Intelligent Systems and Computing (AISC) – Springer, 2015.

[20] S. Jain, K. Asawa, "EMIA: Emotion Model for Intelligent Agent". Journal of Intelligent Systems. Published online. DOI 10.1515/jisys-2014-0071. 2015

[21] S. Planet, and I. Iriondo, "Comparative Study on Feature Selection and Fusion Schemes for Emotion Recognition from Speech", International Journal of Artificial Intelligence and Interactive Multimedia, 1(6), 2012, pp 44-51.

[22] T. Dang,, S. L. Zarshenas, D. Duhaut, "Grace – Generic Robotic Architecture To Create Emotions" . In Proc. Of 11th International Conference on Climbing and Walking Robots and the Suort Technonlogies for Mobile Machines – CLAWAR 2008. Coimbra, Portugal, 2008.

Shikha Jain holds M. Tech in information Technology at GGSIP University, Delhi, India and she is a research scholar at Jaypee Institute of Information Technology, Noida, India where she is working as an assistant professor as well. She has more than 10 years of academic experience. Her interests are focused in the fields of affective computing, intelligent agents, and information retrieval.

Dr. Krishna Asawa presently working with Jaypee Institute of Information Technology (JIIT), Deemed to be University, NOIDA, INDIA in the capacity of Associate Professor. Dr. Krishna awarded Doctor of Philosophy (CSE) in 2002 from Banasthali Vidyapith, Deemed to be University, Banasthali, INDIA. Her area of interest and expertise includes Soft Computing and its applications, Information Security, Knowledge and Data Engineering. Before joining to the JIIT, she worked with National Institute of Technology, Jaipur, INDIA and with Banasthali Vidyapith

Swift vs. Objective-C: A New Programming Language

Cristian González García, Jordán Pascual Espada, B. Cristina Pelayo G-Bustelo, and Juan Manuel Cueva Lovelle,

University of Oviedo, Department of Computer Science

***Abstract* — The appearance of a new programming language gives the necessity to contrast its contribution with the existing programming languages to evaluate the novelties and improvements that the new programming language offers for developers. These kind of studies can show us the efficiency, improvements and useful or uselessness of the new programming languages. Also these studies can show us the good or bad properties of the existing programming languages. For these reasons, these studies allow us to know if the new programming language is offering improvements or relapses.**

In this article, we compare the new programming language of Apple, Swift, with the main programming language of Apple before Swift, Objective-C. We are going to show the differences, characteristics and novelties to verify the words of Apple about Swift. With that we want to answer the next question: Is Swift a new programming language easier, more secure and quicker to develop than Objective-C?

***Keywords* — Object oriented programming, Programming, Functional programming, Programming profession, Software**

I. Introduction

SWIFT is the new programming language created by Apple and it was presented to the public on September 9, 2014 but developers could use it since June 6, 2014. It allows for developing applications for the new version of operating systems of Apple: iOS and OS X [1]. The Apple's intention is to offer a new programming language easier, simpler, more flexible, quicker, funnier and friendly to program than Objective-C [2] to facilitate the applications development for platforms of Apple [1].

Swift was launched to offer an alternative to Objective-C because this has a syntax which barely evolved from it was created and has a great difference with other programming languages that have appeared in the latest years, because these have based on the C++ syntax. For this, Swift is inspired in new programming languages like C++11, C#, F#, Go, Haskell, Java, JavaScript, Python, Ruby, or Scala. Then his syntax is totally different than its predecessor. The Swift's syntax is more simplified because it does not use pointers and includes improvements in its data structures and in its syntax. As we will see, Swift has an easier syntax which helps to developers to have less mistakes and incorporates new functionalities and a new programming paradigm [1].

Mainly, Swift is an object-oriented and imperative programming language as Objective-C but Swift incorporates the functional programming. Some examples of this are the closures, maps and filters.

Due to this facts, it is necessary a study about Swift to check if Swift could be a programming language adapted to the new times and if it could facilitate the application development for platforms of Apple [1].

The remainder of this paper is structured as following: in section III we present the differences among versions. In section IV we discuss the different changes introduced in Swift in front of Objectice-C. Section V explains the new language characteristics. Section VII talks about the novelties that Swift incorporates. In section VI we present the methodology, results and the discussion. Finally, section VII contains the conclusions.

II. Versions

Swift have had different versions with changes in it syntax and functionality since that the first version to developers appeared on June 6, 2014 [1]. At present, Swift is in its third version, Swift 1.2.

The first public version, Swift 1.0 GM, was presented on June 06 2014. It was a Golden Master (GM) version because Apple announced that it will continue adding changes and improvements in the programming language. Swift 1.0 GM presented a lot of changes in its syntax, native libraries and the value type of some function, variable to use the new type "optional" and the syntax of some reserved words like arrays, dictionaries and open range operators.

Swift 1.1, the second version, appeared on October 22, 2014. This update added the "failures initializers", changed some "protocols" and some internal functionalities of Swift.
Swift 1.2 appeared on April 8, 2015. It was a major update. It arrived with the new version of Xcode 6.3. It introduced different improvements in the compiler: the compiler started to create incremental builds; better compilation velocity; improved the error and warnings messages; better stability to avoid the "SourceKit warning". The new language features were: a new reserved word "as!" to clear to readers and developers; the nullability in Objective-C headers; the possibility to export "enumerations" from Swift to Objective-

DOI: 10.9781/ijimai.2015.3310

C with the attribute "@objc"; changes in constants (let); a new native collection: Set.

III. CHANGES RESPECT TO OBJECTIVE-C

Swift and Objective-C use the same compiler, the Low Level Virtual Machine (LLVM). LLVM was created for a student at the University of Illinois in 2000 and it is programmed in C++ [3], [4]. LLVM transforms the Swift source code in optimise native source code for the elected hardware (Mac, iPhone, or iPad) [2].

Swift provides full compatibility with Objective-C and old projects because it allows to use the same libraries, primitive types, control flow and other functions that has Objective-C. However, Swift has various libraries translated to Swift's native code [5].

Furthermore Swift introduces new changes to search to abstract this programming language to the same level that modern programming languages, and in some cases, it obtained more abstraction than them [2], because this is one of the current ways in computer science to help developers when they programme applications and they need to improve their programming [6]–[9]. Also, it has renovated the old syntax of Objective-C. Some examples of this are the classes, protocols, control flows and variables. We are going to explain these similarities and differences in this section.

A. Pre-processor

Swift does not have pre-processor as occurs in C and Objective-C. To achieve the same functionality in Swift, users must use constants instead of the simple macros, namely, define a constant variable and in the case of complex macros, functions [2].

B. Syntax

As in other programming languages (JavaScript, Ruby), Swift allows the optional use of the semicolon character (";") at the end of the line. Besides, Swift uses as access operator the point character (".") like many programming languages instead the square brackets ("[", "]") as Objective-C. This allows more legible code because it has a syntax more similar to the most used programming languages [10], [11].

Furthermore it contains changes in the flow structures' syntax. Now, these must use braces ("{" and "}") to enclose the scope to avoid programming problems. For example, in conditional flow structure ("if"), in some programming languages, in the case that we do not use braces, the first sentence is the sentence that the flow structure will do when the condition will be true and the other sentences will execute in the other cases ("else"). With these changes, Apple wants to do a more legible and easier syntax to developers.

C. Collections

Objective-C has three collections: NSArray, NSSet, and NSDictionary. Initially, Swift only had two collections: Array and Dictionary but in Swift 1.2 Swift added the Set collection. The Swift collections are implemented using structures which differ from the implementation of Objective-C which uses classes.

This difference implicates that in Swift, when these collections are assigned to a constant, variable or they are sent as a function's parameter, Swift creates a copy of them to work with this copy instead the original collection. Meanwhile, Objective-C collections work with the original collection because Objective-C passes references to the original collections instead a copy.

D. Variables

1) Labeled statements

One of the changes in Swift with respect to Objective-C is the functionality of the reserved word "goto". Objective-C allows to use this reserved word to go to any part of the current scope. On the other hand, Swift removed this reserved word and created the "Labeled Statements" [2]. The "Labeled Statements" have a similar functionality as the "goto" but these operate in a smaller scope than the "goto" in Objective-C. Exactly, they have the same scope as other programming languages like C#, Java, and PHP: allow to go to a tag inside a "nested loop" or "switch".

2) Boolean type

Swift, Boolean types have been simplified. Now there only exist the variables "true" and "false". In Objective-C it exists "true"/"false", 0/1, and "TRUE"/"FALSE".

3) Property observers

With Swift has simplified the process to add observers to the variables. To do this, Swift provides two new reserved words: "willset" and "didset" [2]. "willset" is called before the value is assigned to the variable and "didset" is called after the value was assigned to the variable. However, these observers are never called in the first assignation of the instance.

E. Classes and Structures

In Swift, the syntax to create a class or structure is very similar as C++, C#, and Java. Besides, Swift only uses one file (".swift") to define a class, contrary to Objective-C that uses two files (".h" and ".m"). For that, in Swift, you have to define all the class or structure in the same file.

The Fig. 10 contains an example about a definition of a class in Swift. In the first part, it uses a default method, "init", to define the constructor. It keeps the same way and reserved word as Objective-C.

So, to access to the object, Swift still uses the reserved word "self" instead the operator "->". Furthermore, Swift facilitates the access to the properties of the object because it uses the dot operator (".") like C#, Java, JavaScript and Python instead the bracket operators ("[" and "]") as Objective-C. It was possible because Swift removed pointers and it allowed to facilitate the syntax and operators to work with the different properties of the object. For this reason, now, we do not have to allocate a memory block as in Objective-C. Moreover, in Swift we do not have to use the reserved word "new" as in other programming languages when we create an object.

About the inheritance in Swift, it keeps the same way as Objective-C despite now we have to specify when we override the father's method with the reserved word "override" before the child's method and with the reserved word "super" and the dot operator before the father's method when we have to call it.

```
// Class with inheritance
class Child: Person {

    override init(name:String){
        super.init(name: name)
    }

    override func description() -> String {
        return "I'm a child. My name is \(name)."
    }
}

var child = Child(name:"Cris Jr.");
child.name = "Cristian González García Jr.";
var descriptionJr = child.description();
```

Fig. 10. Class example

Apart from this, Swift still allows to avoid the overwriting of a variable, method or class if you include, in the father's part, the reserved word "final". This system uses the same way as C++ and C#.

F. Enumerations

"Enumerations" still have the same functionality. The only changes that they have had are in their syntax. Now, they have a clearer syntax and more similar to a class. Nevertheless, Swift allows to define barcodes, QR codes and raw data with a easier form than Objective-C [2].

G. Extensions

Extensions are used to add new functionalities to an existing class, structure or enumeration to which you cannot access it code. The restriction is that you cannot overwrite the existing functionalities. Extensions are similar to Objective-C categories but without a name [2]. Owing to, for instance, developers can extend the functionality of the String class or others to add new variables or methods.

IV. Language Characteristics

Swift introduced various changes in how to program and it have added new characteristics: it have added changes in variables, it have modified functions and methods to incorporate multiple return and diverse functional programming characteristics. All this will be explained with more details in this section.

A. Variables

Swift is more restrictive than Objective-C because Swift has a strong typing to avoid insecure code [2]. Swift obliges to initialise the variables before their first use. Moreover, you must too specify if the variable is a variable ("var") or a constant ("let") using these reserved word before the name. Besides, Swift checks possible arrays and integer overflow and auto-manages the memory stack using the Automatic Reference Counting (ARC) [2].

Swift allows to the developer to specify explicitly the value type or let the compiler infers the type (Fig. 11), although Swift is strongly typed, for that, when you set a variable in the first time, the compiler assigns it the type and you cannot assign a new value with different type later but you could do an explicit type conversion for changing the type of the value that you want to assign to the type of the variable.

```
// Variable
var variable:String = "I'm a variable"

// Constant
let constant:String = "I'm a constant"

// Static variable
var string:String = "Hello World Swift"

// Inflicted variable
var newString = "Hello World Swift"
```

Fig. 11. Variable examples

B. Functions

Swift allows to send functions as parameter of other functions. Known as Lambda function, Anonymous function, Function literal or Lambda abstraction in Functional Programming. This is one of the characteristics of the Functional programming that Swift has. We show an example in Fig. 12: A – this function receives a function; B it has multiple return; C - one of these return values is a function.

```
func add(array:[Int]) -> Int{
    var aux = 0;
    for i in array{
        aux += i
    }
    return aux
}

func addWithFunction(array:[Int], f:([Int]) -> Int)
    -> (a:Int, b:Int){
    return (f(array), array.count)
}

var array = [Int](0...4)
var result = addWithFunction(array, add)
println(result.a)
println(result.b)
```

A

B

C

Fig. 12. A function that receives a function and returns multiple parameters

C. Classes

Classes incorporate two changes: a new type of constructor, thr "Convenience Initializer" and changes in destructors now known as "Deinitialization". Next, we are going to explain both concepts.

1) Convenience Initializer

The "Convenience Initializer" is an optional constructor that, in case of it exists, it is always called before of the main constructor. Thus, Apple pretends to do the creation of constructors clearly and easier because developers would use different "Convenience Initializer" as alternative constructors [2]. So, normal constructors would have the generic code to the all possible cases. To create a "Convenience Initializers" it is necessary to use using the reserved word "convenience" before the constructor (Fig. 13).

2) Deinitialization

Due that Swift incorporates ARC, for that, destructors are not needed to release memory as occur in C, C++ and Objective-C but Swift incorporates the "Deinitialization" [2].

A "Deinitialization" is a method which is called immediately after an instance is released and it cannot be called explicitly. Using "Deinitialization", developers have a mechanism to do a special clean of different resources or to do actions when the object dies. For example when the object have to work with files. In this case, when the object dies, the program releases the memory that the object had used but the

program cannot close the file because ARC cannot infer this [2]. Then, you can implement a "Deinitialization" to close the file when the object dies. To create a "Deinitialization" you only need to add and implement the method "deinit" as we show in Fig. 13.

```
class Fruit {
    var name: String
    var weight: Int

    init() {
        name = "Unknown"
        weight = 0
    }

    convenience init(name:String) {
        self.init()
        self.name = name
    }

    deinit {
        println("\(name) has disappeared")
    }
}

var fruit = Food()
var fruit1 = Food(name: "Apple")
var fruit2 = Food(weight: 50)
```

Fig. 13. Class example with Convenience Initializers and a Deinit

V. Novelties

In this section we are going to explain the novelties that Swift introduced in relation to Objective-C. The most notable is the "Playground", a sandbox to programming. Also, we are going to explain the novelties in the operators, the new type of variables and some characteristics of the functional programming that Swift introduced as the "Closures". Finally, we are going to talk about the generic which Apple have introduced in Swift, the novelties in the "Switch" flow structure and how to work in Objective-C with Swift and vice versa.

A. Playground

Swift incorporates a new technology in the Integrated Development Environment (IDE) Xcode 6 [12], the "Playground [1], [2]. The "Playground" is a "compiler-in-real-time" which executes the code immediately while the user develops. It provides the value of the execution in different parts of the program (assignments, operations, returns), because the compiler auto-executes itself when it detects a change in the code. Therefore, the "Playground" is a terminal which can execute Swift as other programming languages have like Bash, PHP, Python, and Ruby but with an IDE and some utilities.

This allows to be able to test algorithms in a quick, easy, efficient and isolated way because it gives the possibility to encapsulate the algorithm in a clean sandbox without any relationship to the final application to avoid collateral damages [2].

B. Operators

Swift has all the unary operators (++, --, !), binary operators (+, -, *, /), ternary operators (a:b?c), logical operators (!, &&, ||, true, false), and the assignment of C. Even so, Apple added Range Operators, Overflow Operators and Custom operators. We are going to explain these three operators in the next lines.

1) Range operators

The "Close Range Operator" uses the ellipsis (a…b) as we show in Fig. 14. It defines a range from "a" to "b" and includes both. It is used to iterate on a range where both limits are included.

```
// Closed Range Operator
var closedExample = [0...5]
for index in 1...5 {
    println("\(closedExample)")
}
```

Fig. 14. Closed Range Operator example

On the other hand, the "Half-Open Range Operator" symbol is composed of two dots and the less-than symbol (a..<b) as we can see in Fig. 15. The difference with the "Close Range Operator" is that the "Half-Open Range Operator" does not include the "b" into the range. It defines a range from "a" to "b-1.

```
// Half-Closed Range Operator
var halfClosedExample = [0..<5]
for index in 1..<5 {
    println("\(halfClosedExample)")
}
```

Fig. 15. Half-Closed Range Operator

Using this two operators, we can define in a quick way the creation of a list, array or their iteration.

2) Overflow Operators

In Swift the arithmetic operators (+, -, *, /) do not have "overflow" by default [2]. In case that we want it, we have to add the ampersand sign (&) before the arithmetic operator as we can see in Fig. 16. For example, to apply overflow to the addition, we must use the combination "&+". If we want to have underflow in the subtraction, we have to use "&-". Swift also allows to control the division by zero with the combination "&/" and "&%". For the multiplication we have to use the "&*".

```
var max = Int8.max
//var maxPlus = max + 1 // Error
var maxPlus = max &+ 1

var min = Int8.min
//var minxPlus = max - 1 // Error
var minPlus = max &- 1

//var div = 1 / 0 // Error
var div = 1 &/ 0
var div2 = 1 &% 0

var mul = 1 &* 0
```

Fig. 16. Overflow Operators examples

3) Custom operators

Swift allows to define new operators using the existing arithmetic signs [2]. This is impossible in Objective-C while C++ affords this functionality.

To create a new "Custom Operator", we must to declare its header and do his implementation. There are three ways to do a "Custom Operator": prefix, infix, and postfix (Fig. 17).

```
prefix operator ++++ {}

prefix func ++++ (inout newValue: Int) -> Int {
    newValue += 4
    return newValue
}

var five = 5
var nine = ++++five
```

Fig. 17. Custom Operator

C. Variables

Swift has new data variable types: tuples, "optionals" and "Lazy Stored Property". Some programming languages already incorporated some of these variable types. Then, with this new variable types, Swift offers new ways to work with more flexible and facilities for developers.

1) Tuples

Tuples allow group together multiple values in a unique component. Therefore, one of its uses is to allow the return of multiple values in a function. The Fig. 18 contains an example with tuples.

```
let contact = (13, "Cristian")
var office = contact.0
var name = contact.1
```

Fig. 18. Tuples in Swift

2) Optionals

The "Optional" is a new value type in Swift which neither exist in C nor Objective-C. It is used to assign a type when the value could be of different type or nil [2]. In this way, if a conversion cannot be done, the variable would take a "nil" value. To declare an "optional" variable you must write a question mark (?) after the type. In the example of the Fig. 19, the variable "convertedNumberImplicit" can never be "nil" and in the case that the assignment value would be "nil", the program will break. In the next variable, "convertedNumberExplicit", the type is an Optional Integer. In this case, the value could be "nil".

```
// Optional
let possibleNumber: String = "123"
let convertedNumberImplicit = possibleNumber.toInt()
let convertedNumberExplicit:Int? = possibleNumber.toInt()

let possibleNotNumber = "Hello Swift"
let convertedNotNumberImplicit = possibleNotNumber.toInt()
let convertedNotNumberExplicit:Int? = possibleNotNumber.toInt()
```

Fig. 19. Example that assign value to Optional Variables

This allows to use an optional value in a conditional flow because the "nil" value is the same as a "false" value. However, you can force to read the "optional" value if you write an exclamation mark (!) after the "optional" variable (Fig. 20).

```
if convertedNumberImplicit != nil {
    println("integer value: \(convertedNumberImplicit!)")
} else {
    println("Could not be converted to an integer")
}
```

Fig. 20. Example about how to print Optional Variables

3) Lazy Stored Property

The "Lazy Stored Property" is a class or structure property which value is not calculated until its first use. Until then, it has not value [2]. It uses is similar than other programming languages like C# and Python. In Fig. 21 we create a "Lazy Stored Property" using the declaration "lazy" before the variable.

```
class Student{
    var name: String = "Cristian"
    var surname: String = "González García"
    lazy var contacts = getContacts()
}
```

Fig. 21. Lazy Properties

D. Closures

Swift incorporates the "Closures" as a part of the functional paradigm. The "Closures" allows to evaluate a function in a context which contains one or more variables depending of another context. In Fig. 22, the internal function is executed by the "map function" and it depends on the parameters of the array.

```
var numbers = [20, 19, 7, 12]

numbers.map({
    (number: Int) -> Int in
        let result = 3 * number
        return result
    })
println(numbers)
```

Fig. 22. Applying a Closure using a Map function

Swift allows to create the "Closures" with other simpler ways: implicit return which implicates to not so specify the return; the possibility to use the "Shorthand Argument Names", default arguments created automatically by Swift to work with the parameters sent to the "Closure" without a previous declaration; including just the operator (<, >, +, -,…) in the case that the function only has two parameters and only need to return the result.

E. Generics

Swift adds as a novelty the "Generics", known as "Templates" in other programming languages like C++, and Java. In Objective-C, to have this functionality, you must implement this part of the code in C++.

The incorporation of "Generics" allows to do easier, more flexible, more reusable and more functional with any other data type, avoiding the redundancy of code. This is one of the most powerful functionalities of Swift [2]. As occurs in C++ and Java, the data type used to do the method is generic "<T>". We can see an example in Fig. 23.

```
func genericFunction<T: Comparable>(a: T, b: T) -> T {
    if(a > b){
        return a
    }
    return b
}

genericFunction(5, 3)
genericFunction(3.5, 3.7)
genericFunction("a", "b")
```

Fig. 23. Generic Function

F. Switch

The "Switch" in Swift has improvements in relation with Objective-C [2]. Now, the "break" sentence is optional because the compiler automatically breaks the "switch" when finishes the "case". With this way, Apple pretends to do an easier use and with less programming mistakes. Due to this, it is impossible to have empty "cases" except if you use the reserved word "fallthrough". The reserved word "fallthrough" does that the execution continue to the next "case". The same functionality is in the "Switch" of C#.

As well allows for introducing several verifications in the same "case": you can separate them with commas (,), you can use "Range Operators" or tuples. Also, the "Switch" allows the combination between "Range Operators" and tuples to allow to developers the evaluation of mathematical functions in an easier and quicker way.

Another improvement is the optional clause "where". It allows to add a new additional check in a "case".

G. Objective-C in Swift

Swift allows to incorporate and use Objective-C code in the same Swift program [13]. To insert the Objective-C code, you must insert the Objective-C files (".h" and ".m") in the project and import the Objective-C header in the "Bridging-Header" file which you need to create in the Swift project. This file contains the import all the Objective-C headers (.h) which we will use. After this step, we must connect this file with the configuration through the "Build Settings" configuration. When it is configured, you can work with the Objective-C code using Swift syntax and rules. To import C++ code, you have to create an Objective-C or C wrapper around the C++ code.

H. Swift in Objective-C

To use Swift code in an Objective-C project, the Xcode auto-generates a file when you import the Swift code. This importation allows the layer to access to all the functionalities of the Objective-C code [13]. These files keep the original name but with the additional postfix "-Swift" and the extension ".h". When the importation is done, it just needs to import the Objective-C header files. In this way, you can use the Swift code in an Objective-C project.

VI. Evaluation and Discussion

In this section we are going to explain in detail the different process of evaluation and then we show the results. Firstly, we will describe the used methodology. After that, we will show the results and discuss them.

A. Methodology

We have evaluated the implementation of the same code in each different programming language: Objective-C and Swift. We have chosen an example of an XML parser, exactly, using the Foundation XMLParser. In our case, we implemented a main method and override two methods: "didStartElement" and "foundCharacters". The first one, "didStartElement", is called when the parser needs to process a new XML node. The second one, "foundCharacters", is called when the parser find text into a node.

After that, we evaluate the code using two ways. Firstly, we analysed and compared the syntax. Secondly, we did a quantitative analysis based on source code: we counted the lines with code; the words, reserved words and the numbers of "Switch cases"; the characters that developers need to write that code.

B. Results

In this section, we are going to explain the two ways that we used to compare Objective-C with Swift. Firstly, we are going to talk about the syntax analysis where we have implement three methods in both programming languages. Later, we are going to present the quantitative analysis where we compare both in base on their lines, words, and characters using the same three methods of the syntax analysis.

1) Syntax Analysis

Firstly, we analysed the "beginParsing" method in both programming languages (Fig. 24 and Fig. 25). As we can see, the Objective-C implementation uses a pointer. Furthermore, when it initialise the parser, it needs to allocate memory.

```
-(void)beginParsing:(NSURL *)xmlURL
{
    parser = [[NSXMLParser alloc] initWithContentsOfURL:xmlURL];
    [parser setDelegate:self];

    [parser setShouldProcessNamespaces:NO];
    [parser setShouldReportNamespacePrefixes:NO];
    [parser setShouldResolveExternalEntities:NO];

    [parser parse]
}
```

Fig. 24. "BeginParsing" method in Objective-C

```
func beginParsing(xmlUrl :NSURL)
{
    parser = NSXMLParser(contentsOfURL: xmlUrl)

    parser.delegate = self
    parser.shouldProcessNamespaces = false
    parser.shouldReportNamespacePrefixes = false
    parser.shouldResolveExternalEntities = false

    parser.parse()
}
```

Fig. 25. "BeginParsing" method in Swift

Secondly, in the "didStartElement" method (Figure 17 and Figure 18) we can observe similarities with the previous method. Objective-C uses pointers and need to allocate memory. Besides, you have to create a copy to avoid the use of the original. Another difference between Objective-C and Swift is the obligatory use of the "at sign" character (@) before a "string".

```
- (void)parser:(NSXMLParser *)parser didStartElement:(NSString *)elementName namespaceURI:(NSString *)
    namespaceURI qualifiedName:(NSString *)qualifiedName attributes:(NSDictionary *)attributeDict
{
    element = [[elementName copy] stringByTrimmingCharactersInSet:[NSCharacterSet
        whitespaceAndNewlineCharacterSet]];

    if ([element isEqualToString: @"views"]) {
        elements = [[NSMutableDictionary alloc] init];
        views = [[Views alloc] init];
    }
}
```

Fig.26. “didStartElement” method in Objective-C

```
func parser(parser: NSXMLParser!,didStartElement elementName: String!, namespaceURI: String!,
    qualifiedName : String!, attributes attributeDict: NSDictionary!)
{
    element = elementName.stringByTrimmingCharactersInSet(NSCharacterSet.
        whitespaceAndNewlineCharacterSet())

    if (element as NSString).isEqualToString("views") {
        elements = NSMutableDictionary.alloc()
        elements = [:]
        views = Views()
    }
}
```

Fig. 27. "didStartElement" method in Swift

Finally, we show the implementation in Objective-C (Fig. 28) and Swift (Fig. 29) of the method “foundCharacters”. In this case, we can see the same differences again. In addition, Swift needs less characters because it omitted the use of the reserved word “break”.

```
- (void)parser:(NSXMLParser *)parser foundCharacters:(NSString *)string
{
    int number = [string intValue];

    switch(number){
        case 0:
            [button setText: cad];
            break;
        case 1:
            [input setText: cad];
            break;
        default:
            NSLog(@"Default case");
            break;
    }
}
```

Fig. 28. Implementation of the "foundCharacters" method in Objective-C

```
func parser(parser: NSXMLParser!, foundCharacters string: String!)
{
    var number: Int = string.toInt()

    switch(number){
        case 0:
            button!.setText(cad)
        case 1:
            input!.setText(cad)
    default:
        println("Default case")
    }
}
```

Fig. 29. Implementation of the "foundCharacters" method in Swift

1) Quantitative Analysis

Now, we show the quantitative analysis about the three methods which were implemented. First of all, we show the comparison between Objective-C (OC) and Swift (S) about the lines with code in the Table 1. As we can see, the difference depends on the method but it is irrelevant. For this reason, we cannot say that with Swift you could write much less code lines.

TABLE 1
COMPARISON: LINES WITH CODE

Lines	OC	S	Difference	%
beginParsing	6	6	0	0
didStartElement	6	7	-1	14,28
foundCharacters	12	9	3	-33,33

SECONDLY, IN THE

Table 2 we show the comparison about the numbers of words that we had needed to implement the methods. In this case, we used less words in the case of Swift but the difference is insignificant. Swift is a few less verbose than Objective-C as we explained in the previous sections.

TABLE 2
COMPARISON: WORDS

Words	OC	S	Difference	%
beginParsing	23	22	1	-4,54
didStartElement	34	31	3	-9,67
foundCharacters	30	28	2	-7,14

Finally, we measured the number of characters of each method in both programming languages. In this case, we did not count the whitespaces. In the Table 3 we can see that Swift needs around 11% less characters than Objective-C. This is because Swift has changed the access operators, does not use pointers and removed the obligation of the use the semicolon and some reserved words.

TABLE 3
COMPARISON: CHARACTERS WITHOUT SPACES

Characteres	OC	S	Difference	%
beginParsing	256	226	30	-13,27
didStartElement	401	360	41	-11,38
foundCharacters	211	187	24	-12,83

VII. CONCLUSION

Apple has achieved to create a modern programming language with same of the best functionalities of other programming languages like C#, Java, JavaScript, PHP,

Python, Ruby, and Scala, among others. Furthermore, Swift incorporates the functional programming as other programming languages have adopted to provide more possibilities to developers. In addition, Swift adds new possibilities for doing easier and more effective the applications development owing to the fact that: they have done change in the syntax, classes, variables, functions, operators, and data structures; they have improved the "Switch"; they have removed of the pointers.

Furthermore, we can see that using Swift, developers need less characters to program the same code because Swift has simplified the syntax but they need the same numbers of lines and words to program.

Because of these reasons, it seems a wise decision of Apple the creation of a new programming language with more abstraction level than Objective-C to facilitate the application development but with all power of other current programming languages and the elimination of the obsolete Objective-C syntax. Moreover, this abstraction allows to make less mistakes and be more comfortable to developers without a flexibility loss and maintenance the retro-compatibility with previous code.

In conclusion, Apple has created a programming language with the necessary abstraction and functionalities of this age and, in certain cases, they have improved the current functionalities as they have done with the "Switch". Besides, they have created a programming language to improve their ecosystem due to the fact that Swift is not a competitor of Objective-C nor its evolution, Swift is a programming language ready to coexist with Objective-C and to give another possibility to develop to developers.

ACKNOWLEDGMENT

This work was performed by the "Ingeniería Dirigida por Modelos MDE-RG" research group at the University of Oviedo under Contract No. FUO-EM-086-14 of the research project "Proyecto Visio". Project partially-financed by Zed Worldwide S.A.

REFERENCES

[1] Apple Inc., "Swift," https://developer.apple.com/swift/, 2015. [Online]. Available: https://developer.apple.com/swift/. [Accessed: 17-Apr-2015].
[2] Apple Inc., The Swift Programming Language. 2014.
[3] "LLVM," http://llvm.org/, 2000. [Online]. Available: http://llvm.org/.
[4] C. A. Lattner, "LLVM : An Infrastructure for Multi-Stage Optimization," University of Illinois, 2002.
[5] @adamjleonard, @thinkclay, and @cesar_devers, "Swift Toolbox," http://www.swifttoolbox.io/, 2014. [Online]. Available: http://www.swifttoolbox.io/. [Accessed: 17-Apr-2015].
[6] E. González, H. Fernández, and V. Díaz, "General purpose MDE tools," Int. J. Interact. Multimed. Artif. Intell., vol. 1, pp. 72–75, 2008.
[7] E. R. Núñez-Valdez, O. Sanjuan-Martinez, C. P. G. Bustelo, J. M. C. Lovelle, and G. Infante-Hernandez, "Gade4all: Developing Multi-platform Videogames based on Domain Specific Languages and Model Driven Engineering," Int. J. Interact. Multimed. Artif. Intell., vol. 2, no. Regular Issue, pp. 33–42, 2013.
[8] R. Gonzalez-Crespo, S. R. Aguilar, R. F. 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," Int. J. Interact. Multimed. Artif. Intell., vol. 1, no. 7, pp. 63–69, 2012.
[9] C. G. García, C. P. García-Bustelo, J. P. Espada, and G. Cueva-Fernandez, "Midgar: Generation of heterogeneous objects interconnecting applications. A Domain Specific Language proposal for Internet of Things scenarios," Comput. Networks, vol. 64, no. C, pp. 143–158, 2014.
[10] TIOBE Software BV, "TIOBE Index," http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html, 2014. [Online]. Available: http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html. [Accessed: 15-May-2015].
[11] C. Zapponi, "GitHub," 2014. [Online]. Available: http://githut.info/. [Accessed: 15-May-2015].
[12] Apple Inc., "Xcode." 2015.
[13] Apple Inc., Using Swift with Cocoa and Objective-C. 2014.

Cristian González García is a Technical Engineering in Computer Systems and M.S. in Web Engineering from School of Computer Engineering of Oviedo in 2011 and 2013 (University of Oviedo, Spain). Currently, he is a Ph.D. candidate in Computers Science. His research interests are in the field of the Internet of Things, Web Engineering, Mobile Devices and Modeling Software with DSL, and MDE.

Jordán Pascual Espada is a Research scientist at Computer Science Department of the University of Oviedo. Ph.D. from the University of Oviedo in Computer Engineering. His research interests include the Internet of Things, exploration of new applications and associated human computer interaction issues in ubiquitous computing and emerging technologies, particularly mobile and Web.

B. Cristina Pelayo G-Bustelo is a Lecturer in the Computer Science Department of the University of Oviedo. Ph.D. from the University of Oviedo in Computer Engineering. Her research interests include Object-Oriented technology, Web Engineering, eGovernment, Modeling Software with BPM, DSL and MDA.

Juan Manuel Cueva Lovelle is a Mining Engineer from Oviedo Mining Engineers Technical School in 1983 (Oviedo University, Spain). Ph. D. from Madrid Polytechnic University, Spain (1990). From 1985 he is Professor at the Languages and Computers Systems Area in Oviedo University (Spain). ACM and IEEE voting member. His research interests include Object-Oriented technology, Language Processors, Human-Computer Interface, Web Engineering, Modeling Software with BPM, DSL and MDA.

A Quantitative Justification to Dynamic Partial Replication of Web Contents through an Agent Architecture

Enrique Torres[1], José Daniel García[2], Oscar Sanjuan[3], Luis Joyanes[4] and Rubén González Crespo[3]

[1]*Madrid Council IT Deparment, Madrid, Spain*
[2]*Universidad Carlos III de Madrid, Leganés, Madrid, Spain,*
[3]*Universidad Internacional de La Rioja (UNIR), La Rioja, Spain,*
[4]*Universidad Pontificia de Salamanca, Madrid, Spain*

***Abstract* — The most usual solution to improve the performance of a Web server is based on building a distributed architecture, where the Web server is offered from a set of nodes. The most widely distributed architecture is based on Web clusters including a Web switch. The Web switch is responsible for deciding which site's node must attend which request. When deciding where elements are stored the classical solution was to fully replicate all contents in every server node. However, partial replication may require a fraction of storage while offering the same level of reliability. In this paper we report a solution based on dynamic partial replication where the number of replicas for each file and its management is handled by an agent architecture. We compare our solution with full replication and with static partial replication both in terms of storage capacity consumption and service time. Our results show that our proposed solution provides equivalent performance with a better use of disk storage capacity.**

***Keywords* — Distributed computing, Computer network performance, Network servers, Cooperative systems.**

I. Introduction

A Web site typically consists of a set of elements or resources, where each of them can be of a certain type (HTML page, image, video, file download, music, etc.). A page consists of a primary element that refers to a series of secondary elements (included in the page).

The Web site receives requests from clients, where each request for a web page typically consists of multiple requests to the server, one for each object included in that page. The client establishes a connection to the server for each of these requests, and receives the response for this connection [1], although some optimizations are possible to reduce the number of connections.

The most usual solution to improve the performance of a Web server is based on building a distributed architecture, where the Web service is offered from a set of nodes [2], acting as a logical single server and consequently giving a single server image.

There are a variety of solutions allowing the construction of distributed Web servers [3]. Those solutions are based either on total replication of contents (all contents are replicated in every server nodes) or on the distribution of contents (contents are distributed so that each element is in a single server node).

Between these two alternatives (total replication and total distribution), a third alternative can be found, namely partial replication, where a certain number of copies is performed for each element. This third alternative [4] [5] needs to determine a priori the number of copies required, which does not always respond to the real needs of the system or to the needs evolution over time.

An evolution of this third alternative dynamically adapts the number of stored replicas according to actual needs and modifies this number when change is needed [6].

This paper presents a quantitative evaluation performed on a prototype that follows this model in which the access time and required storage space is quantified.

The rest of this paper is structured as follows: Section II describes the architecture used in the distributed web server, Section III explains algorithms developed to allow dynamic replication, a prototype is presented in Section IV, results from evaluation are shown in Section V, finally conclusions are drawn in Section VI.

II. A Dynamic Partial Replication Architecture Implementation by Agents

The most widely distributed Web system architecture is based on clusters [1]. In this architecture (Fig. 1) a distribution node between clients and servers is used: the Web switch. The Web switch is receiving all requests to the visible IP address through a request distribution algorithm, and decides which server node should process which request [3]

The use of an additional service network (Fig. 2) is a modification from the base architecture of a Web-based system cluster that can improve performance of adaptive allocation of contents, for the redistribution of those contents without affecting the main network of the cluster [7].

DOI: 10.9781/ijimai.2015.3311

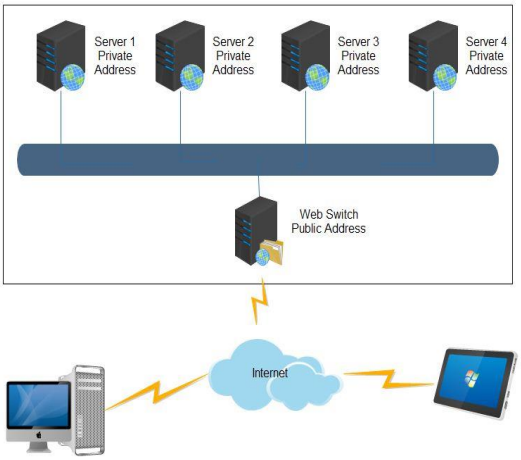


Fig. 1. Cluster Web based General Architecture

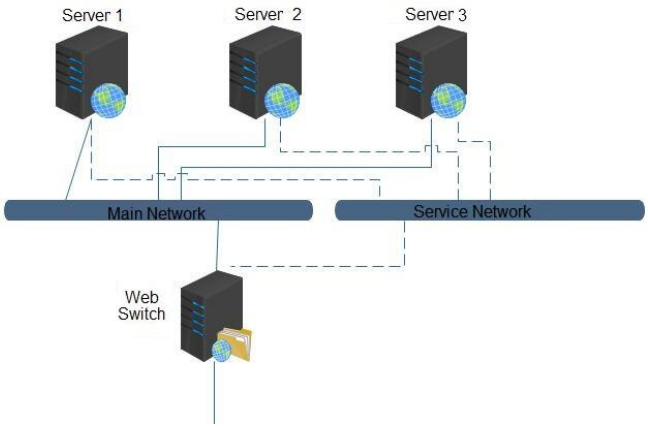


Fig. 2. Cluster Web based General Architecture with an Additional Service Network

Development through intelligent agents simplifies decision making for each component, as each agent behaves in an autonomous way and interacts with the rest of agents [8].

The needed implementation to develop a Web site with partial replication needs to perform activities in the Web switch as well as in server nodes.

- The Web switch must contain all the necessary logic to allocate requests to servers (Request Distributor Agent), and to account them in order to modify the popularity and to determine whether the popularity change implies a change in replication (Replica Control Agent). This machine must also contain the module allowing the administrator to add or remove items (External Control Agent). The Web switch scheme is presented in Fig. 3.
- The server node must include the required logic to obtain an item when the Web switch decides that there should be an additional replica on that node making a request to other server node (Content Control Agent), and the component performing requests resolution by accessing disk (Disk Access Agent). The server node scheme is presented in Fig. 4.

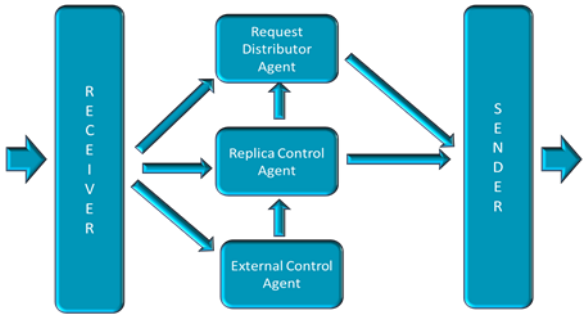


Fig. 3. Agents in a Web switch

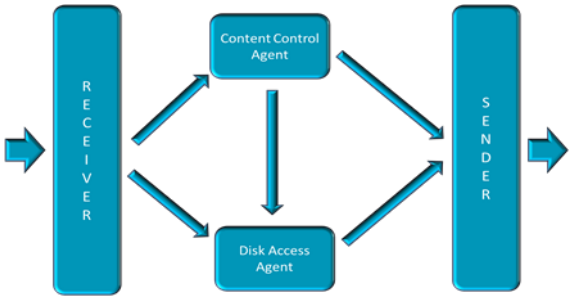


Fig. 4. Agents in a Server Node

To build the previous model, it is necessary to specify the algorithm used by each of the components, on the understanding that the receiver and transmitter modules only transfer requests reaching each system

III. Dynamic Replication Algorithms

Having established the architecture to be used, it is also necessary to explain the proposed dynamic replication algorithm, which will be placed in the above mentioned replica module. This algorithm tries to dynamically optimize the number of copies depending on the number of requests it receives to each file. Consequently, it will increase the number of replicas of those files which are in high demand to meet actual requests while it will reduce the number of copies of those files which are in low demand in order to free up storage space in nodes [9].

The dynamic replication faces three issues: number of copies for each file, choosing the nodes where the copies are stored, and when the algorithm is executed.

- **Number of copies**: The algorithm assigns all files once, causing the free storage of nodes to decrease by a certain percentage. This percentage decreases the probability of files to be requested and, while this value is greater than 0, the system includes a copy of that file. This will give the total number of copies of file, so we compare this number with the number of copies already in the system and add or remove copies depending on whether the new number is higher or lower [10]. The algorithm for obtaining the number of copies is shown in Algorithm 1.

```
assign TRUE to insert
while insert is equal to TRUE
    assign FALSE to insert
    for all i = 0 to FileNumber
        if FileProbability i > 0
            decrease size i of FreeSpace
            increase CopyNumber
            assign TRUE to insert
    for all i = 0 to FileNumber
        decrease (1-FreeSpace)/TotalSpace
            of FileProbability i
```

Algorithm 1: Obtaining the number of copies.

- **Storage node**: When assigning copies of files to different nodes, it is important to note that nodes in the proposed system have a finite storage space and this aspect must be taken into account [11]. In addition, cloned nodes should be avoided as much as possible, and an appropriate algorithm should be used. A widely used solution addressing the same problem [7] is the backpack greedy algorithm with prior sorting of replicas by size.
- **Activation time**: The system should consider to increase the number of replicas of a file when quality of service is compromised [6]. That is, when the response time of a file exceeds a preset time, the situation is called timing failure. The maximum response time for each file and the probability that the response time is met will be established by the administrator. The latter value is considered because it is admissible that a small percentage of timing failures does not compromise the quality of service and therefore while failures are below the threshold the number of copies will not be increased. When the request ratio that meets the established response time is below the established probability, the algorithm is activated to optimize the number of copies of the files. Moreover, every time a new file is added, the number of copies must be recomputed for all the files, using the probability given by the Pareto distribution for the new file.

IV. Model Built

To quickly create a model to perform the evaluation, we decided to build a simulation of the real situation.

To build model we used the OMNeT ++ with the INET framework. This tool has already been used to create a large number of projects, such as developing a full suite of TCP/IP at the Karlsruhe University [12], a framework for computer architectures simulation [13], a file storage model for distributed systems [14], a simulation model for IEEE 802.15.4 [15], or make a performance analysis of a handover level 2 in IPv6 mobile networks [16] among many others.

Three VLANs are required for this model, the first routes the client requests to the Web Switch, the second links the Web Switch with the server nodes, and the third one is used as internal service network. A simplified diagram is shown in Fig. 5.

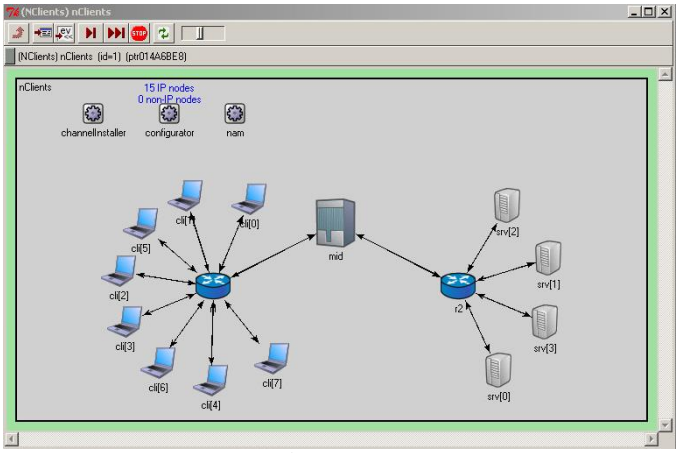


Fig. 5. Model Architecture simplified to 8 clients

Behavioral patterns of a web server based on experimental data were presented in several papers [17] [18]. In this paper, an adaptation of this model has been made to the case of a Web cluster.

A Web site consists of a set of elements or resources, where each may be of a certain type (HTML page, image, video, file download, music, etc.). Set E (see Equation 1) can be defined as the set of all elements that make up the Web site.

$$E = \{e_1, e_2, \ldots, e_N\} \quad (1)$$

Each element of set E can be a primary element or a secondary element. Thus, the E set can be expressed as the union of set E_p, primary elements, and set E_s, secondary elements (see Equation 2).

$$\begin{aligned} E &= E^p \cup E^s \\ E^p &= \{e_1^p, e_2^p, \ldots, e_n^p\} \\ E^s &= \{e_1^s, e_2^s, \ldots, e_m^s\} \end{aligned} \quad (2)$$

The number of secondary elements included per each primary element can be modeled by a Pareto distribution [19] (see Equation 3).

$$f(x) = \frac{\alpha k^\alpha}{x^{\alpha+1}} x \geq k \quad (3)$$

A web page can be defined as a pair consisting of a primary element and a set of secondary elements, and this allows define the W set as all the site's web pages (see Equation 4).

$$\begin{aligned} w_i &= \{e_i^p, e_{i1}^s, \ldots, e_{in}^s\} \\ W &= \{w_1, w_2, \ldots, W_m\} \end{aligned} \quad (4)$$

The size of the elements is modeled [19] by a lognormal distribution (see Equation 5).

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \quad (5)$$

Experimental studies have established values of α = 2.43 and k = 1 to the number of secondary elements [19], as well as values μ = 9.537 and σ = 1.318 for size of primary elements and μ = 8.215 and σ = 1, 46 for the size of secondary elements [20] [21].

Activity of a client is determined by a sequence of sessions and downtime between sessions (*Downtime*) as show in Fig. 6.

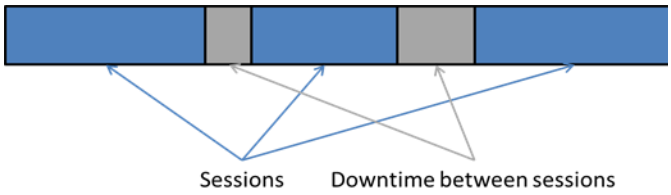


Fig. 6. Customer's activity along the time

Downtime can be modeled by a Pareto distribution [19]. Experimental studies [22] use values α = 1.4 and k = 20.

During a session, a client visits a set of Web pages (*Requests per session*), starting with the entry page to the website. Before moving to the next page, it evaluates the contents of the current page for a given time (*Inactivity time*) as shown in Fig. 7.

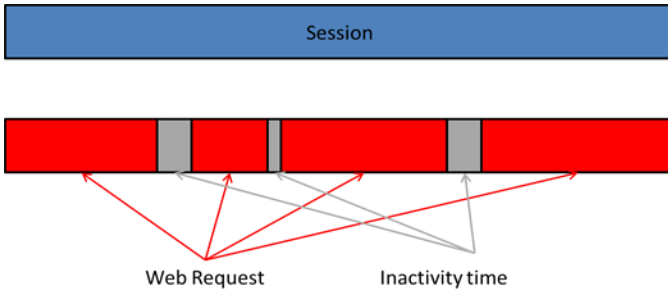


Fig. 7. Session activity

The *Requests per session* can be modeled quite accurately by an inverse Gaussian distribution [23] (see Equation 6) with experimental values of μ = 3.86 and λ = 9.46 [20] [21].

$$f(x) = \sqrt{\frac{\lambda}{2\pi x^3}} e^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}} \quad (6)$$

Several studies have modeled the downtime by a Pareto distribution with experimental values k = 1 and α = 1.4 [20] [21].

After receiving the response to the primary element, the client scans the content (*Scan time*) before generating multiple request and multiple connections, up to the maximum indicated by the *Degree of concurrency*, one for each secondary file.

The *Scan time* is modeled by a lognormal distribution with experimental results of μ = 360.4 and σ = 106.5 [18].

The *Degree of concurrency* has been modeled as a constant which sets the number of connections.

In the server the different elements involved in an file access [23] are included to compute the access time to each of the stored files. The file requests arrive to the File Manager that passes it to the I/O Manager.

The File Manager receives file requests to be read/write and queries the File Table. In this article was used a UNIX-like file system structure [24], with some simplifications that do not affect the access time. First the Index Node Table is looked up to obtain the disk blocks that must be accessed and if the file uses indexing blocks, Fig. 8, the system keeps the access order to know the addresses of blocks.

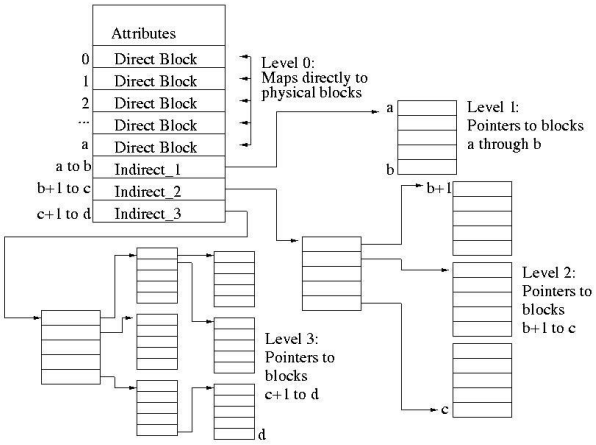


Fig. 8. Index Node Table in UNIX

A call to the I/O Manager is made for each block, which computes the physical address, scheduling disk accesses according to these addresses and computes access time based on the current position of the head and the physical parameters of each of the disks [24]. It is possible to set both the rotational speed, the speed of movement of the heads and the transfer rate.

Most of the time used in the resolution of an HTTP request corresponds to the operation of disk access for the file associated with the request, which depends largely on the size of the file.

V. Evaluation Results

We compared different options for content distribution, in order to evaluate whether replica allocation policy affects performance.

In all alternatives it has been simulated the behavior of 200 clients making requests on a web cluster system for 7 days of simulated time.

The alternatives evaluated were:

TREP	Total Replication
SPREP	Static Partial Replication
DPREP	Dynamic Partial Replication

Iteration over the servers for each of the files is used as requests allocation policy. First server that contains the file requested is selected both **SPREP** and **DPREP**.

The first performance metric used is the *HTTP Request Service Time*, which corresponds to the time between the time

when the client sends a request for a file and the time when it receives the response. Table 1 shows the average results of the three alternatives when the number of servers is increased.

TABLE I
SIMULATION RESULTS FOR THE AVERAGE SERVICE TIME OF A WEB SITE WHEN THE NUMBER OF SERVERS IS INCREASED

Number of servers	TREP	SPREP	DPREP
4 servers	1,82601405	1,33961388	1,34021504
5 servers	1,57350203	1,59601244	1,58789328
6 servers	1,62635074	1,76683025	1,75924017
7 servers	1,45933568	1,53470846	1,50389218
8 servers	2,01486739	1,99620641	1,98940237
9 servers	1,28018552	1,24920451	1,31952067
10 servers	1,4591151	1,4672851	1,4619481

To determine whether the difference in Average Service Time of a Web Site is significant, an analysis of variance test has been performed, with the results shown in Table 2. The test was performed for a value of α = 0, 05

TABLE II
VARIANCE TEST RESULTS FOR THE AVERAGE SERVICE TIME OF A WEB SITE WITH A=0,05

F	0,0110429
F Critical Value	3,55455714
Probability	0,98902453

Fig. 9 shows graphically the above average values. It can be easily seen that the Average Service Times are similar.

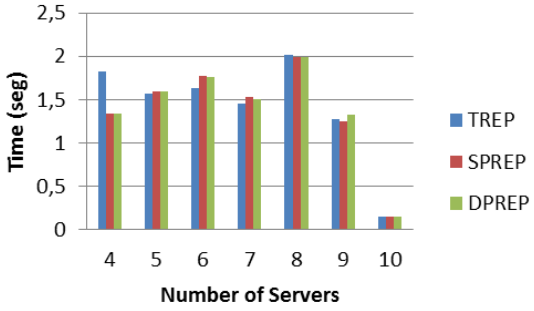


Fig. 9. Average Service Time of a Web Site

Another metric to evaluate the evolution of the storage space is the number of files stored in each server. The average results were shown in Table 3.

TABLE III
FILES' COPIES STORED IN EACH SERVER

Number of servers	TREP	SPREP	DPREP
4 servers	10400	8361,00	8157,00
5 servers	10400	7953,20	7684,60
6 servers	10400	7681,33	6427,00
7 servers	10400	7683,43	5818,90
8 servers	10400	7683,52	5853,50
9 servers	10400	7683,52	5859,85
10 servers	10400	7614,88	5876,13

If we consider the distribution of files is the same for primary and secondary files, and according to several studies [20] [21] it is possible to calculate the number of secondary files associated to a primary as it follows a Pareto distribution with α = 2, 43 k = 1, (see Equation 7).

$$E(X) = \frac{\alpha\, k}{\alpha - 1} = 1{,}6993 \qquad (7)$$

We can deduce that the function that determines the number of primary and secondary files based on the total number of files stored on a server is given by the expressions in Equation 8.

$$primaries\ (n) = \frac{n}{2{,}69} \qquad secondaries(n) = \frac{1{,}69\, n}{2{,}69} \qquad (8)$$

And since the distribution following the primary and secondary files [17] is known [18], we can compute the average of these values (see Equation 9).

$$E_{prim}(X) = e^{7{,}63 + \frac{1{,}001^2}{2}} = 3398{,}1977$$
$$E_{secon}(X) = e^{8{,}215 + \frac{1{,}46^2}{2}} = 10730{,}0125 \qquad (9)$$

Using the number of files expressed in Table 3 we can compute the average storage required in each server, as show in Table 4.

TABLE IV
AVERAGE STORAGE NEEDED IN EACH SERVER

Number of servers	TREP	SPREP	DPREP
4 servers	79,39 GB	63,82 GB	62,27 GB
5 servers	79,39 GB	60,71 GB	58,66 GB
6 servers	79,39 GB	58,63 GB	49,06 GB
7 servers	79,39 GB	58,65 GB	44,42 GB
8 servers	79,39 GB	58,65 GB	44,68 GB
9 servers	79,39 GB	58,65 GB	44,73 GB
10 servers	79,39 GB	58,13 GB	44,85 GB

Fig. 10 shows graphically the average storage needed in each server in the different evaluations performed.

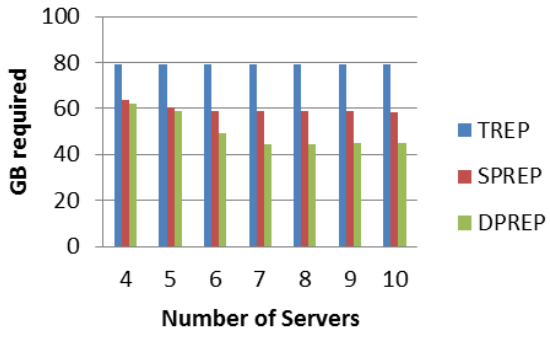


Fig. 10. Average Storage needed in each server

If we also consider the number of servers in each of the simulations, it is possible to compute the average total number of files stored in the Web site, as shown in Table 5.

TABLE V
AVERAGE TOTAL NUMBER OF FILES IN THE WEB SITE

Number of servers	TREP	SPREP	DPREP
4 servers	41600	33444	32628
5 servers	52000	39766	38423
6 servers	62400	46088	38562
7 servers	72800	53784	40732
8 servers	83200	61468	46828
9 servers	93600	69152	52739
10 servers	104000	76149	58761

Allowing us to establish the percentage of space needed in the different solutions versus TREP, as shown in Table 6.

TABLE VI
USAGE RATE VERSUS TREP

Number of servers	TREP	SPREP	DPREP
4 servers	100%	80,39 %	78,43 %
5 servers	100%	76,47 %	73,89 %
6 servers	100%	73,86 %	61,80 %
7 servers	100%	73,88 %	55,95 %
8 servers	100%	73,88 %	56,28 %
9 servers	100%	73,88 %	56,34 %
10 servers	100%	73,22 %	56,50 %

These rates are graphically shown in Figure 11.

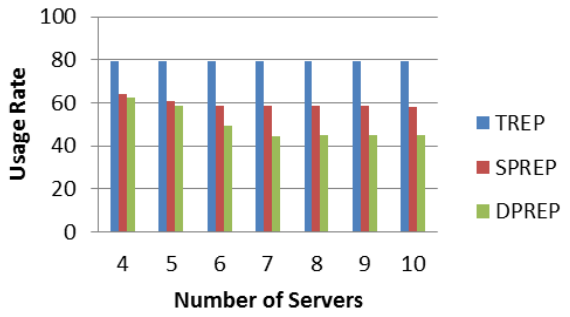


Fig. 11. Usage Rate versus TREP

VI. CONCLUSIONS

This paper presented a web cluster architecture allowing partial replication of website contents with dynamic adaptation of the number of replicas.

A prototype was developed to test the architecture. The proposed structure is intended to provide fault tolerance, simplicity and distribution taking advantage of the use of an agent architecture.

The results of this prototype are presented in this article, evaluating both performance and storage capacity.

We have seen no evidence to reject the hypothesis that the Request Service Time does not depend on content allocation policy. Even if we admit that there are differences, they never exceed 0.03%.

It has been shown that the necessary storage volume is much greater in the case of full replication that in the other cases. Dynamic partial replication is placed at the other extreme, and storage capacity required is the least of all strategies studied.

REFERENCES

[1] T. Berners-Lee, R. Fielding and H. Frystyk, "Hypertext Transfer Protocol - HTTP/1.0. RFC 1945", Internet Engineering Task Force, Mayo 1996.

[2] V. Cardellini, E. Casalicchio, M. Colajanni, and P.S. Yu, "The state of the art in locally distributed Web-server systems", *ACM Computing Surveys*, vol. 34(2):263–311, 2002.

[3] T. Schroeder, S. Goddard and B. Ramamurthy, "Scalable web server clustering technologies", *IEEE Network*, vol. 14(3), pp. 38–45, Mayo 2000.

[4] García, J.D., Carretero, J., García, F., Singh, D.E. y Fernández, J.: "A Highly Available Cluster of Web Servers with Increased Storage Capacity", *XVII Jornadas de Paralelismo*, pp. 109-114, Albacete, Septiembre 2006.

[5] Garcia, J.D., Carretero, J., Garcia, J., Sánchez, L.M. y Garcia, F.: "A Web Cluster Architectural Proposal for Balancing Storage Capacity and Reliability by using Partial Replication", *International Journal of Computer Systems Science and Engineering*, vol. 28(3):191-202, CRL Publishing Ltd, Reino Unido, Mayo 2013.

[6] Torres, E., Sanjuan, O., Joyanes, L., García, J.D. y González, R.: An Architecture For Management Of Distributed And Redundant Web Storage With Intelligent Agent Systems And Emerging Techniques. En IEEE Latin America Transactions, vol. 6(6):524-528, IEEE Xplore, Octubre 2008.

[7] J.D. Garcia, "Propuestas Arquitectónicas para servidores Web distribuidos con réplicas parciales", Ph.D. dissertation, Computer Science Dept., Universidad Carlos III de Madrid, Junio 2005.

[8] Gutierrez, C.: "An Analysis Architecture for Communications in Multi-agent Systems", *International Journal of Artificial Intelligence and Interactive Multimedia (IJIMAI)*, vol 2(1):65-72, 2013.

[9] Torres, E., Sanjuan, O., Joyanes, L., García, J.D. y Pelayo B.C.: "A Multi-Agent based Proposal for the Management of Distributed and Redundant Storage". *Proceedings of the 2008 International Conference on Artificial Intelligence (ICAI 2008)*, pp. 566-571, Las Vegas, Nevada, USA, CSREA Press, Julio 2008.

[10] Torres, E., Sanjuan, O., Joyanes, L., García, J.D., González, R. y Ríos, S.: "Management of Distributed and Redundant Storage in High Demand Web Servers for Heterogeneous Networks Access by Agents", *International Symposium on Distributed Computing and Artificial Intelligence (DCAI 2008)*, vol. 50, pp. 123-131, Advances in Soft Computing, Springer, Enero 2009.

[11] Torres, E., Sanjuan, O., Joyanes, L., García, J.D. y González, R.: "Arquitectura Para La Gestión De Almacenamiento Web Distribuido Y Redundante Mediante Sistemas De Agentes Inteligentes Y Técnicas Emergentes". *6th International workshop on practical applications on agents and multi-agent systems (IWPAAMS2007)*, pp. 319-328, Salamanca, Noviembre 2007.

[12] Kaage, U., Kahmann V., Jondral F: "An OMNeT++ TCP Model", *Proceedings of the 15th European Simulation Multiconference (ESM 2001)*, Praga, Junio 2001.

[13] Núñez, A., Fernández, J., Carretero, J., García, J.D. y Prada, L.: "SIMCAN: A SIMulator Framework for Computer Architectures and Networks!, *First International Workshop on OMNeT++*, pp. 8, Marsella, Francia, Marzo 2008.

[14] Núñez, A., Fernández, J., Carretero, J., García, J.D. y Prada, L.: "New Techniques for Modelling File Data Distribution on Storage Nodes", *41st Annual Simulation Symposium*, pp. 175–182, Ottawa, Canada, Abril 2008.

[15] Kirsche, M. y Schnurbusch, M.: "A new IEEE802.15.4 Simulation Model for OMNeT++/INET", *First OMNeT++ Community Summit*, Hamburg, September 2014.

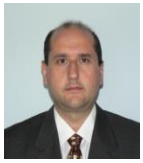
[16] Çetin, G., Çetin, A. y Özkaraca, O.: "The analisys of Layer-2 Handover performance for mobile IPV6 using OMNeT++ Simulation Tool", *Mugla Journal of Science and Technology*, vol. 1(1):34-38, 2015.

[17] Mah, B.A.: "An empirical model of HTTP network traffic", *Proceedings of the Conference on Computer Communications (INFOCOM'97)*, vol. 2, pp. 592-600, Kobe, Japón, 1997.

[18] Choi, H.K. y John O.L.: "A behavioral model of web traffic", *Proceedings Seventh International Conference on Network Protocols (ICNP'99)*, pp. 327-334, Toronto, Ontario, Canada, IEEE, Octubre 1999.

[19] Barford, P. y Crovella, M.: "Generating representative web workloads for network and server performance evaluation", *Performance Evaluation Review*, vol. 26(1):151-160, Junio 1998.

[20] V. Cardellini, M. Colajanni, and P.S. Yu, "Geographic load balancing for scalable distributed Web Systems", *Proceedings of the 8th International Symposium on Modeling, Analisys and Simulation of Computer and Telecommunication Systems (MASCOTS'00)*, pp. 20-27, IEEE, San Francisco, CA, USA, Agosto 2002.
[21] V. Cardellini, "Request redirection algorithms for distributed web systems", *IEEE Transactions on Parallel and Distributed Systems*, vol 14(4), pp. 355-368, Abril 2003.
[22] Wallerich, J.: "Design and implementation of a www workload generator for the ns-2 network simulator", *http://www.net.t-labs.tu-berlin.de/~joerg/nsweb/doku/*, Noviembre 2001. [consulta 1 de Abril de 2015].
[23] Stalings, W.: "Operating Systems: Internals and Design Principles (8th Ediдtion)", Ed. Prentice Hall, 2015.
[24] Carretero, J., Miguel, P., García, F. y Pérez, F.: Sistemas Operativos: una visión aplicada 2ª Edición, Ed. McGraw-Hill 2007.



D. Enrique Torres has a degree in Computer Science from the Polytechnic University of Madrid. He has been Assistant Professor at the Faculty of Computer Science from the Pontifical University of Salamanca. Ha has also been an Adjunct Professor at University Carlos III of Madrid. He is currently a systems engineer at the Madrid Council.

Dr. José Daniel García has a degree in Computer Science from the Polytechnic University of Madrid and a Computer Science PhD from Universidad Carlos III de Madrid.He is Associate Professor in Computer Science Department from University Carlos III of Madrid, where he previously served as Assistant Professor. Previously he was Lecturer in the Computer Science Faculty from Pontifical University of Salamanca. Dr. García has worked as consultant and software engineer in projects for several international companies like FCC, Siemens, DMR Consulting, Telefónica, or ING Bank. His main research interests including parallel and distributed systems, programming languages and programming models for applications improvement.

Dr. Oscar Sanjuan has a degree in Computer Science from the Pontifical University of Salamanca, where he also earned his PhD in Computer Science Engineering, and he is PhD in Computer Science from Oviedo University. He has been Area Director of Software Engineering at the Pontifical University of Salamanca, lecturer and researcher at the University of Oviedo and Assistant Professor at the University Carlos III of Madrid. Currently he is Engineering Director at ElasticBox Inc.

Dr. Luis Joyanes Aguilar has a degree in Physics Science from the Complutense University of Madrid and he has a degree in Military Higher Education from General Military Academy of Zaragoza. He earned his PhD in Computer Science from Oviedo University and his PhD in Sociology from Pontifical University of Salamanca. Also, he is Honorary Doctor from the Private University Antenor Orrego of Trujillo in Perú. He is Professor of Languages and Computer Systems from Pontifical University of Salamanca and he is member of Knowledge Management Committee of AEC (Spanish Quality Association) and AENOR (Spanish Agency for Standardization).

Dr. Rubén González has a degree in Computer Science from the Pontifical University of Salamanca, where he earned his PhD in Computer Science Engineering. He has been Area Director of Operating Systems at the Pontifical University of Salamanca and Graduate Director from Higher School of Engineering and Architecture from Pontifical University of Salamanca. Currently, he is Director of the Higher School of Engineering from International University of La Rioja and Director of the AENOR Chair in Certification and Quality and Technology Standards.

Semi-Automated Correction Tools for Mathematics-Based Exercises in MOOC Environments

Alberto Corbí and Daniel Burgos

Universidad Internacional de La Rioja

***Abstract* — Massive Open Online Courses (MOOCs) allow the participation of hundreds of students who are interested in a wide range of areas. Given the huge attainable enrollment rate, it is almost impossible to suggest complex homework to students and have it carefully corrected and reviewed by a tutor or assistant professor. In this paper, we present a software framework that aims at assisting teachers in MOOCs during correction tasks related to exercises in mathematics and topics with some degree of mathematical content. In this spirit, our proposal might suit not only maths, but also physics and technical subjects. As a test experience, we apply it to 300+ physics homework bulletins from 80+ students. Results show our solution can prove very useful in guiding assistant teachers during correction shifts and is able to mitigate the time devoted to this type of activities.**

Keyword s— assignments, semiautomated correction, maths, physics, framework.

I. INTRODUCTION

MOOCs and online campuses nowadays represent an observable reality when it comes to self-education [5]. Together with *OpenCourseWare* platforms, they are definitively impacting our current TEL scene. Even in MOOC environments, students are usually required to carry out some homework. Nevertheless, these homework bulletins are hardly ever supervised by a tutor or a teacher. Quite the opposite, the students themselves are required to self-correct and self-assess their exercises based on correction grids, templates and answer keys. Peer reviewing also takes place, as we will discuss in section II. Fully automated quizzes are also commonly displayed and correction is normally done by the MOOC and/or e-learning platform.

Technical documents from the STEM fields (Science, Technology, Engineering and Mathematics) increase document richness with many sorts of structured objects: mathematical and chemical formulae, diagrams, tables and relations, etc. These additions usually carry essential information that complements the texts the student has to read. At first sight, homework assignments related to these disciplines are good candidates for automated correction processes. However, many teachers are interested not only in the accuracy of the result but also in the correctness of the resolution process, which might turn out to be as important as –or sometimes even more important than– the final outcome itself. Corrections performed by a human (a teacher/assistant) can also add value to the teacher's view on how his/her students learn and progress. The teacher's feedback on a correction sheet always entails a unique opportunity to improve the learner's knowledge and build a more robust awareness on the matter they are currently working on.

Exercises in physics deepen this reviewing philosophy and student-teacher interaction. Keeping an organized and coherent resolution flow is as relevant to the understanding of the underlying physical phenomena as the final output itself.

Besides, in physics, results can belong to a broad spectrum of mathematical natures and entities, ranging from simple and isolated numbers or scalars (e = 2.7182), vectors ($\vec{v} = 3\vec{i} + \vec{j}$), signed quantities (-*k*) and physical units (3.3 kΩ), to name a few that might appear on a basic physics course. In addition, slightly different numbers, notations and/or symbols can represent exactly the same correct result and account for the same reality. For instance, $\left|\vec{B}\right| = 5\,\mathrm{N/Am}$ and $\|\mathbf{B}\| = 4.99\ \mathrm{T}$ can both be labeled as correct and the student should receive a positive score/comment. If such minor discrepancies could be detected, an automated system might be able to send back an explicit recommendation as [26], for example, does. In the same sense, and as a last example, all of the following expressions have the exact same meaning: partial differentiation of function *f* with respect to an independent variable *x*:

$$f_x,\ f'_x,\ d_x f,\ \partial_x f,\ \nabla_x f,\ \partial f/\partial x,\ D_x f$$

Finally, students attending physics courses in online institutions and/or MOOCS come from very different backgrounds and behavior is easily altered over time, as described by [1]. The *human touch* in the reviewing process has always proven to be the key to success, independently of the academic environment: online, formal, higher education, etc.

All this being said, in MOOC environments, the amount of homework bulletins to be reviewed, and the substantial tutoring effort that takes place if every exercise from every

DOI: 10.9781/ijimai.2015.3312

student is manually revised, can reach disproportionate levels. One of the goals of our project is concerned with assisting teachers during the correction phase. This target is achieved by pre-classifying student bulletins as ready to be *teacher-reviewed* or not. In the latter case, an automated message can be issued to the student, who can re-edit his/her own document before reissuing it to the teacher, for a second time. Of course, this assistant tool would heavily depend on the type of subject and content to be analyzed. In this paper, we focus on assisting teachers in online campuses and MOOCS when reviewing homework related to mathematical content.

II. Overview of the Current Status of MOOCs, Online Education and Student Assignment Management

MOOCs face nowadays a number of challenges: accreditation management, credit recognition, monetization implementation and content and methodology quality assurance. Among them, methodology quality becomes the foundation from which the other four are built. MOOCs are taking over the long-tradition role of Open Educational Resources. Some MOOCs also combine face-to-face strategies with online learning and even merge formal and informal settings. In addition, MOOCs highlight the current need for basic and specific competence acquisition, as a complement to the current courses, very much focused on personal interests and continuing education. They are also turning out the ultimate tool to fight against the lack of access to teaching resources (disadvantaged individuals, regions and countries).

MOOC platforms require support for teachers and tutors, based on their needs, skills, and teaching context. One of these has to do with grading essays and activities. Since MOOCs seeks the enrollment of hundreds or thousands of students, the evaluation becomes a real challenge. At present, some MOOCs rely on peer-assessment and counseling. Peer-to-peer seems significant and useful, so there is, at first sight, no need for a replacement. However, a complementary evaluation resource would be welcome by the educational community.

There are some approaches for automatic or semi-automatic assessment, like ontology networks [23], where the conceptualization of the domain model becomes the cornerstone to categorize and shapes the results properly. Another strategy involves the temporal hiring of additional teachers as *graders*, so they can act as complement to those professors officially assigned to the course. In addition, a detailed comment and assessment on the submitted final activity might not be compulsory, as long as the learner does not require a formal accreditation. This strategy scales down the number of assessments to those learners who actually send a formal/official request. At Universidad Internacional de La Rioja (UNIR), there is a prototype implemented, and under testing phase: A4Learning [30]. This tool is integrated into the Sakai LMS, and retrieves behavioral and academic information from users, so that they can be compared with previous records. Out of this comparison, the tool makes an estimate on every student on how his/her progress will be, based on similar profiles. In doing so, the professor gets a detailed analysis of every learner, 1 by 1, and clustered by similarity. With A4Learning, the teacher can analyze the student current status, anticipate potential academic future, and react in consequence. There is another early prototype, AppMOOC, which will retrieve basic requirements to grade activities, so that, when the professor gets an essay, a previous checking mechanism guarantees that the work fulfills these minimum information and/or structure. These two prototypes, A4Learning and AppMOOC, will be implemented along the next academic year at a larger scale, with the clear objective of supporting teachers on their functions as evaluators and feedback providers, big mid-size and large-size groups of learners, worldwide. The research work described in this paper is in intimate relation with the aforementioned projects.

III. Towards an Automated Helper System for Maths and Technical Student Homework Preliminary Screening

We have designed a special workflow and protocol that automatically analyses student assignments and checks whether they contain coherent mathematical information related to specific fields. This set of tools also takes into account equivalent expressions, exemplified in section I.

In order to check for this coherence, simple –but also highly configurable and easily editable– content-checking rules designed by the teacher are submitted to the correction engine. Then, for every exercise in the student digital notebook, mathematical expressions are semantically compared with the correction template submitted by the teacher. A more detailed review of the practical implementation is tackled below.

Of course, designing such a protocol is no easy task and has required working with state-of-the-art mathematical language-processing techniques and mathematics representation standards, also reviewed below.

A. State-of-the-Art Language Processing in Mathematics

Despite the fact that linguistic analysis of scientific documents is currently seen as an interesting line of research, the current work in the field is still limited. Mathematical literature represents a rather isolated linguistic niche embodying its own challenges. We can identify a significant contrast between this linguistic realm and, for instance, the domain of medical/healthcare research publications that have been studied by many scientific groups in recent years. Two of the current main issues that make mathematical texts challenging to work with are:

- Natural language –expressing complex symbolism– and mathematical representation are usually mixed and hosted in the same document.
- Almost a complete absence of accurately labeled linguistic compilations.

Indeed, state-of-the-art analyses largely try to bypass these problems by restricting their scope to *well-formed* sections of mathematical text and reports, as in the *controlled* approach reviewed below.

The first challenge of the recognition process is the recovery of the so-called *layout tree* [9] of the mathematical

expression. The next step involves creating operator trees. These trees are data structures that hold the logical relationships within an equation, as opposed to its horizontal and vertical links. The structure of the mathematical expression can then be made computationally transparent, which is necessary for any practical application involving a mathematics recognition process, like the one we are introducing in this paper. The layout tree also carries a burden of uncertainty in its correctness, which adds to the difficulty of establishing the expression's logical structure.

A holistic and detailed analysis of the processes of extracting and retrieving mathematical expressions and mathematics recognition has already been carried out by [28].

We will now review some lines of enquiry that have recently attracted interest in the research community around math semantics and language processing.

1) Controlled Natural Language

In this approach, a restrained natural language for mathematics is incrementally built [12]. With it, we are then capable of supporting a sufficient subset of natural language elements that would allow an author to write math expressions in a simple way but also be limited enough to allow unambiguous interpretation. Its primary goal is building formalized libraries of mathematical content, focusing on establishing pipelines over a narrow subset of language. Next, a systematic and careful widening takes place. Current projects implementing this view are:

- FMathL [21] described in `mat.univie.ac.at/~neum/FMathL.html`
- MathLang [13]
- MathNat [12]
- Naproche [3], [6], available at `naproche.net`

2) Natural Mathematical Discourse

The opposite 2of the controlled approach is to try to model the original language of real scientific documents [6, 29]. Consistent work in the area has been developed by [27] and [11], as well as by [4]. The corpus used for this work is based on the arXMLiv archiving project of scientific documents [24]. arXMLiv is hosted at the Cornell arXiv (`arxiv.org`) which contains one of the largest collections of scientific literature on the planet. Unfortunately, its texts are in the TeX/LaTeX format, which makes it rather useless for knowledge analysis engines, even though LaTeX can be considered a *de facto* global standard of typesetting. The goal of the project described in [10] is to translate all these documents to a common and agreed XML scheme, which can then serve as a basis for revealing math-related semantics.

B. Computer Representation of Mathematical Content with LaTeXML

LaTeXML [7] uses a context-free grammar to establish the logical structure of a document with mathematical content. It can then be exported to Content MathML and OpenMath [2]. Content MathML (also referred as MathML v3 from the W3C consortium and described in `w3c.org/TR/MathML3`) uses just a few attributes and focuses on the meaning of the expression rather than its graphical layout. The `<apply>` element, for instance, represents the application of a function. Its first child element is the function itself and its operands and/or parameters are the remaining child elements.

Thanks to Content MathML and Open Math, digital libraries can be transformed into a more useful XML representation and be made more compliant with a mathematical knowledge-management approach. Two large-scale examples are arXMLiv and EuDML [22]. Only the first of those examples uses LaTeXML. The main challenges in this conversion step, in the case of arXMLiv, come from the fact that it is poorly knowledge-based, with minor exceptions in the form of clues provided via some infrequent and almost random in-line LaTeX annotations. It is then mandatory to infer additional semantics on all document levels. Fortunately, LaTeXML has proven to be extremely efficient at this task.

Consider the example in Fig. 1. There we have the standard mathematical notation –a simple equation of the form *f(x) = y*–, its Content MathML representation and, finally, the terms we extracted for indexing. Any mathematical construct can be represented in a similar way.

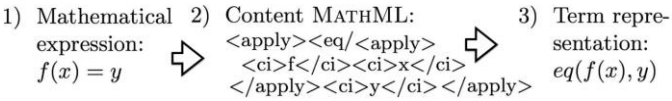


Fig. 1. From plain mathematical expressions to Content MathML.

LaTeXML also defines a conversion process and a set of tools that allow any plain LaTeX document to be translated [7]. LaTeXML can even work in daemon mode, which allows the deployment of server-centric conversion platforms [8] like the well-known *ltxMojo*, available at `latexml.mathweb.org`.

Once a mathematical text has been retyped as LaTeXML, search queries can take place. This topic is discussed in the following section.

C. The MathWebSearch Project

MathWebSearch [17], developed at the KWARC group (`kwarc.info`), processes XML-based content mathematics. Currently, the system supports MathML, OpenMath and LaTeXML (and any other document type that has been appropriately converted). It operates by computing an index term for each of the mathematical elements of a given XML document. Queries on this index are also expressed in a XML schema, reviewed below.

The MathWebSearch engine is used in our framework to analyse student-submitted mathematics assignments. On one hand, each student document is converted to Content MathML and indexed. On the other hand, a teacher's set of well-organized binary tests is coded as a variant of Content MathML –MathMLQ–. If all tests deliver a positive result, the assignment is flaged as *to be reviewed* by the teacher.

Finally, as MathWebSearch operates with terms, heuristics and semantics, it can understand a wider range of similar mathematical expressions. This ensures that the issues described in the introduction will hardly ever take place. Our engine is very tolerant to small variations of the same

mathematical expression. In other words, we are able to understand that $\sqrt{x}$ and $x^{1/2}$ have the same mathematical meaning and discern that 4.5 kJ is different from 4.5 Kj (the Joule energy unit in physics must always be capitalized, while the *kilo-* multiple should remain in lowercase). In this manner, the student is free to express him/herself with mathematical and syntactical independence. At the same time, the teacher is also able to demand exquisite precision, if so desired.

Algorithm 1 Example of a MATHMLQ query.

```
<mws:query
xmlns:mws="http://www.mathweb.org/mws/ns"
xmlns:m="http://www.w3.org/1998/Math/MathML"
limitmin="0" answsize="30" totalreq="yes"
output="json">
 <mws:expr>
  <m:apply>
   <m:csymbol cd="ambiguous">
    superscript
   </m:csymbol>
   <mws:qvar>q</mws:qvar>
   <m:cn type="integer">2</m:cn>
  </m:apply>
 </mws:expr>
</mws:query>
```

D. The MathWebSearch Query Language

MathWebSearch makes use of a content-oriented query language called MathMLQ. It is XML-based rather than being a genuine query language by itself. More detailed information on the syntax can be found in [18]. An example of application can be read in algorithm 1. The query described there is able to identify both the square of a function or a variable (x^2 or $[f(x)]^2$).

Apart from describing queries using the MathMLQ syntax just introduced, more simple instances can be expressed using the plain LaTeX math toolbox and syntax. This code can be then converted to MathMLQ. This conversion takes place with the tool `latexmlc`, presented in [16], which can also establish relations between LaTeX and a variety of office documents (WML from MS Word, ODT from Open Office, etc.) In this simplified LaTeX syntax, variables are labeled with the question mark symbol (`?`). For instance, the following expression:

```
latexmlc --address = latexml.mathweb.org/
convert --preload=mws.sty --whatsin=math –
whatsout = math --cmml 'literal:\sqrt{?c}^2'
```

Would produce the same XML output as the one displayed in algorithm 1.

E. Summary of Implementation

We now summarize the skeleton of our software implementation, which is graphically represented in Fig. 2. Students submit their homework in a variety of formats (Microsoft Office Word, OpenOffice, OpenDocument, Portable Document Format, LaTeX and LyX, etc.). Disciplines related to theoretical fields, such as mathematics, physics and computer science, almost exclusively use LaTeX. On the other hand, more applied fields of research, like life sciences, chemistry and engineering, usually typeset on the so-called *office suites*. Moreover, depending on the discipline, each institution has its own focus and teachers expect homework to be edited using a specific software instance.

For this reason, our system tries to, in the first phase, convert each document type to a unified LaTeX representation. This is not always possible due to technical reasons (converter segmentation fault, faulty output, etc.). Several third party tools (both open source and commercial) exist and operate with greater or lesser degrees of success. Writer to LaTeX (`writer2latex.sf.net`) and Word to LaTeX (`wordtolatex.com`) are some examples. LyX has the advantage of being able to perform a clean LaTeX export [14].

A better tool to translate between LaTeX and traditional office formats is the `latexmlc` introduced above, which has been developed in recent years by the KWARC group. Finally, the tool that has recently been attracting significant focus in the computer language research community is Pandoc, described in [20] and [19]. Pandoc can convert documents in markdown, HTML, LaTeX, MediaWiki markup, TWiki markup, Microsoft Word docx and EPUB (among others) to other formats, such as DocBook, Adobe InDesign, LaTeX, PDF and many others, through the application of external drivers written in the Lua computer language.

Anecdotally, recent efforts are even trying to directly translate mathematical handwritten expressions to LaTeX. A nice summary can be found in [25] and an example of such an application can be tested online thanks to Detexify [15], available at `detexify.kirelabs.org`.

As a next step, the LaTeX source is parsed and transformed to LaTeXML, which already contains the necessary knowledge companion information to be harvested by MathWebSearch. On the other side, the teacher pulls a list with *N* wildcard expressions to the classification platform. Finally, an instance of MathWebSearch performs these *N* searches on each homework document and screens which of them provides some degree of equivalence. Our platform is responsible for filtering teachers' templates and student homework in a coordinated fashion.

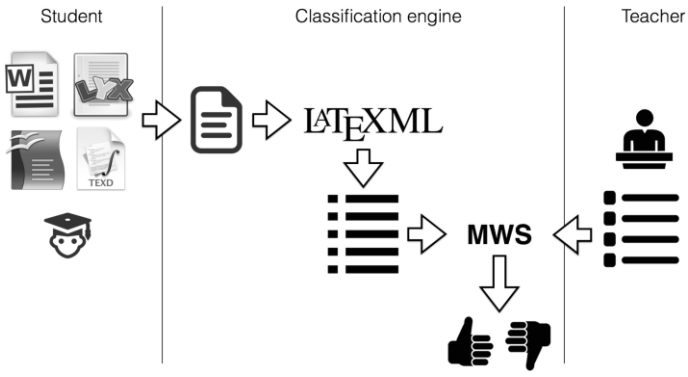


Fig. 2. Overview of our platform. Students submit their homework and a conversion process to LaTeXML takes place. On the other side, teachers feed the system templates with mandatory mathematical expressions.

IV. Practical Experience with Physics Students' Homework Bulletins

As a proof of concept, we have carried out a practical experience with 300+ homework assignments from 50+ students enrolled in a basic Physics course in the degree of Computer Science at the School of Engineering at Universidad International de La Rioja (UNIR, `ingenieria.unir.net`).

We have configured our classification engine based on MathWebSearch together with teachers' templates in order to pre-distribute assignments, before they are finally delivered to the teacher/assistant for an in-depth (and manual) conventional correction phase.

A. Experimental Setup

The online campus platform deployed at UNIR is an instance of the Apereo Sakai CLE. Students submit their homework to this platform digitally, using the assignments tool. Usually, documents are formatted using Microsoft Word®, WML or OpenOffice ODT, though some students have used LaTeX or LyX for their submissions. A very small percentage of students submitted bulletins in other office suite formats, such as Apple Pages® or Microsoft PowerPoint®, which were easily translatable to WML or ODT.

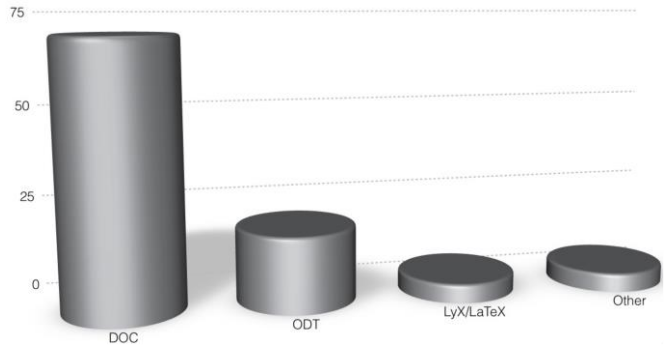


Fig. 3. Percentage of document types used by students.

The rate of conversion success to LaTeX and LaTeXML from this range of commonly available office suites is summarized in Table I. After running each of the conversion tools, further refinement can take place if the source office documents are pre- or post-manually processed.

TABLE I.
Conversion level of success from office documents submitted by students to LaTeX and LaTeXML

Source *office* format	% of success	after manual edition
WML (recent MS Word ver.)	75 %	90 %
.doc (legacy MS Word ver.)	70%	90%
ODT (OpenOffice)	80 %	95 %
Other office suites	50 %	70 %
LyX	100 %	100 %

The conversion tool most used in our setup, given its success ratio, was Pandoc, as described above. Fig. 4 shows a real example of the result of the conversion of a MS Word-submitted homework file to its LaTeX twin. PDF output (from LaTeX) is also shown as a proof of the fidelity of the file-translation process.

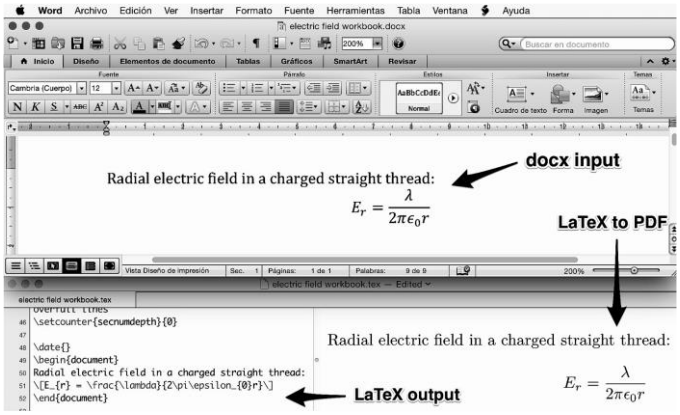


Fig. 4. Example of the conversion process performed with Pandoc.

Fig. 5 shows the ratio of success in the process of translating to LaTeX, of some of the file-conversion tools that are mentioned above and were used in this project.

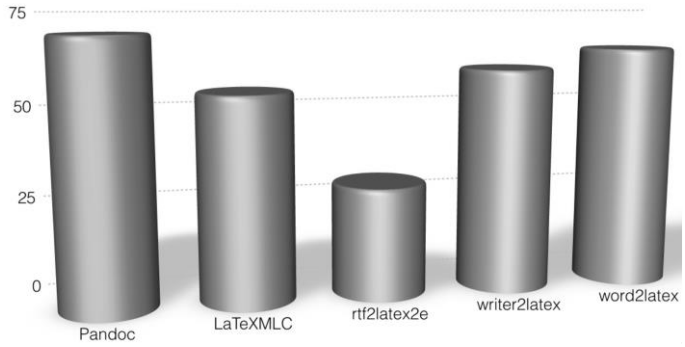


Fig. 5. Rates of success for some of the converter tools (to LaTeX).

B. Methodology

The physics course mentioned above, as it is part of the Computer Engineering degree's curriculum, is mainly based on areas related to electromagnetism. Most required homework exercises should include at least some of the mathematical expressions appearing in table II –depending on the specific topic being studied– in order to be considered suitable for further analysis by the teacher and manually assigned a score. This mathematical content has been agreed with the academic staff. The corresponding set of simplified queries (introduced above) has also been defined and has been made available to the system.

In Table III, there is another example of how our implementation can also handle more complex formulae, for instance those related to quantum theory and thermodynamics, which could prove useful in a Physics MSc.

Our solution has been tested offline (no real feedback has been sent to students or teachers) with pre-existing homework bulletins from an already concluded semester. A batch process, similar to that described in Fig. 2, has been implemented and executed.

Besides taking into account specific mathematical content related to the topic *electromagnetism*, we have also established a special and separate realm devoted only to pure mathematical transversal correctness. This means that our solution can separately test for the exactitude of common mathematical statements, like the ones listed in Table IV

TABLE II.
SOME MATHEMATICAL EXPRESSIONS RELATED TO THE TOPIC ELECTROMAGNETISM TO BE TESTED.

Electric field by an infinite plane (perpendicular)	$E_x = 2\pi k\sigma$
Electrostatic force (vector form)	$\vec{F} = k\frac{\lvert q_1 q_2\rvert}{r^2}\hat{r}$
Charge linear density	$dq{=}\lambda dx$
Capacitor equation	$C = \epsilon\frac{S}{d}$
Electric field by a charged ring	$E_x = \frac{kQx}{(x^2+R^2)^{\frac{3}{2}}}$
Electric field by a charged circle	$E_x = 2\pi k\sigma\left(1 - \frac{x}{\sqrt{x^2+R^2}}\right)$
Electric field by a charged thread	$E_y = k\lambda\frac{L}{y\sqrt{\frac{1}{4}L^2+y^2}}$
Electric potential by a set of isolated charges	$V = \sum_i \frac{kq_i}{r_{i0}}$
Relation between potential and electric field	$dV(r) = -\mathbf{E}(r)\cdot dr$
Electric field outside a charged a spherical cavity (radial component)	$E_r = \frac{1}{4\pi\epsilon_0}\frac{Q}{r^2}$

TABLE III.
EXTENDED PHYSICS-RELATED EXAMPLES.

Heat diffusion law	$\frac{dQ_x}{dt} = -kA\frac{\partial T}{\partial x}$
Entropy	$S = \sum_i p_i \ln p_i$
Tunnel effect	$T = \frac{e^{-2\int_{x_1}^{x_2} dx\sqrt{\frac{2m}{\hbar^2}(V(x)-E)}}}{\left(1+\frac{1}{4}e^{-2\int_{x_1}^{x_2} dx\sqrt{\frac{2m}{\hbar^2}(V(x)-E)}}\right)^2}$
Momentum by a photon	$p_f = \frac{h\nu}{c}$
Gyro dynamics	$\omega_p = \frac{MgD}{\mathbf{I}_s\omega_s}$
Charged particle in accelerator	$\mathbf{F} = q(\mathbf{E}+\mathbf{v}\times\mathbf{B})$

TABLE IV.
TRANSVERSAL MATH EXPRESSIONS.

Two-component vector definition	$\vec{V} = (V_x, V_y)$
Vector sum	$\vec{V_1} + \vec{V_2} = (V_{1_x} + V_{2_x}, V_{1_y} + V_{2_y})$
Vector length	$\lvert\vec{V}\rvert = \sqrt{V_x^2 + V_y^2}$
Vector dot-product	$\vec{V_1}\cdot\vec{V_2} = V_{1_x}\cdot V_{2_x} + V_{1_y}\cdot V_{2_y}$
Scalar-vector product	$N\cdot\vec{V} = N\cdot(V_x, V_y) = (N\cdot V_x, N\cdot V_y)$

With this external test, our system allows teachers to filter bulletins based only on pure mathematical fidelity, ignoring topic-specific inaccuracies or errors

V. RESULTS AND DISCUSSION

After running a batch process with the 300+ homework bulletins and specific rule sets, results show that around 63% of the documents that could be safely converted to LaTeX satisfied the formulae template requirements (both for the topic electromagnetism and for the transversal one related to mathematics). Of these homework assignments, 78% were given a positive score by the teacher at the moment of the reviewing process. The remaining 22% of documents that were classified as incorrect, though encapsulating the required mathematical expressions, contained inaccuracies and/or were poorly developed by the student.

VI. CONCLUSIONS

Our simplified and relatively quick set-up proves that semi-automated correction processes may represent an acceptable compromise between the pure self-assessment approach –typically present in MOOCS and courses with a large enrolment rate– and the more conventional scenario in which the teacher manually reviews assignments for each student.

ACKNOWLEDGMENT

This work has been partially funded by UNIR Research Funding Programme 2013-2015, under the TELSOCK Research Group, the UNESCO Chair on eLearning and the ICDE Chair in Open Educational Resources.

REFERENCES

[1] John M Aiken, Shih-Yin Lin, Scott S Douglas, Edwin F Greco, Brian D Thoms, Michael F Schatz, and Marcos D Caballero. The initial state of students taking an introductory physics mooc. arXiv preprint arXiv:1307.2533, 2013.

[2] Ron Ausbrooks, Stephen Buswell, David Carlisle, Giorgi Chavchanidze, Stéphane Dalmas, Stan Devitt, Angel Diaz, Sam Dooley, Roger Hunter, Patrick Ion, Michael Kohlhase, Azzeddine Lazrek, Paul Libbrecht, Bruce Miller, Robert Miner, Murray Sargent, Bruce Smith, Neil Soiffer, Robert Sutor, and Stephen Watt. Mathematical Markup Language (MathML) version 3.0. W3C Recommendation, World Wide Web Consortium, 2010.

[3] Marcos Cramer, Peter Koepke, and Bernhard Schröder. Parsing and disambiguation of symbolic mathematics in the Naproche system. Proceedings of the 18th Calculemus and 10th International Conference on Intelligent Computer Mathematics, 2011.

[4] J.H. Davenport, W.M. Farmer, F. Rabe, and J. Urban. Intelligent Computer Mathematics: 18th Symposium, Calculemus, and 10th International Conference on Artificial intelligence, 2011.

[5] Luis de-la Fuente-Valentín, Aurora Carrasco, Kinga Konya, and Daniel Burgos. Emerging technologies landscape on education. a review. International Journal of Interactive Multimedia and Artificial Intelligence, IJIMAI, 2(3), pp.55, 2013.

[6] Mohan Ganesalingam. The Language of Mathematics. PhD thesis, Cambridge University, 2009.

[7] Deyan Ginev. LaTeXML: A LATEX to XML converter, arXMLiv branch. Web Manual. 2011.

[8] Deyan Ginev. The LaTeXml daemon: Editable math for the collaborative web.

[9] Deyan Ginev. The structure of mathematical expressions. PhD thesis, Master Thesis, Jacobs University Bremen, 2011.

[10] Deyan Ginev, Constantin Jucovschi, Stefan Anca, Mihai Grigore, Catalin David, and Michael Kohlhase. An architecture for linguistic and semantic analysis on the arXMLiv corpus. Applications of Semantic Technologies Workshop, 2009.

[11] Helmut Horacek and Magdalena Wolska. Handling errors in mathematical formulas, Lecture Notes in Computer Science, 2006.

[12] Muhammad Humayoun and Christophe Raffalli. Mathnat: mathematical text in a controlled natural language. 9th International Conference on Mathematical Knowledge Management, 2010

[13] Fairouz Kamareddine. Motivations for mathlang, T-O-U Workshop, 2005.

[14] David Kastrup. Revisiting WYSIWYG paradigms for authoring LaTeX. Proceedings of the 2002 Annual Meeting, TUGboat, 2002.

[15] D. Kirsch. Detexify: Erkennung handgemalter LaTeX-symbole. PhD thesis, Westfälische Wilhelms-Universität Münster, 2010.

[16] Lukas Kohlhase and Michael Kohlhase. System description: A semantics-aware LaTeX-to-office converter. Intelligent Computer Mathematics, 2014

[17] Michael Kohlhase, Bogdan A. Matican, and Corneliu-Claudiu Prodescu. Mathwebsearch 0.5: Scaling an open formula search engine. Lecture Notes in Computer Science, 2012.

[18] Michael Kohlhase and Ioan Sucan. A search engine for mathematical formulae, Artificial Intelligence and Symbolic Computation, 2006.
[19] Jacco Krijnen, Doaitse Swierstra, and Marcos O.Viera: Towards an extensible Pandoc, Lecture Notes in Computer Science, 2014.
[20] John MacFarlane. Pandoc user's guide. http://johnmacfarlane.net/pandoc/README.html, 2013.
[21] Arnold Neumaier and Peter Schodl. A framework for representing and processing arbitrary mathematics. KEOD, 2010.
[22] Jiri Rakosnik and Radoslav Pavlov. European digital mathematics library. Digital Presentation and Preservation of Cultural and Scientific Heritage, 2013.
[23] Lucía Romero, Milagros Gutiérrez, and María Laura Caliusco. Conceptualizing the e-learning assessment domain using an ontology network. International Journal of Interactive Multimedia and Artificial Intelligence, IJIMAI, 1(6), pp. 20, 2012.
[24] Heinrich Stamerjohanns and Michael Kohlhase. Transforming the arxiv to XML. In Intelligent Computer Mathematics, 2008.
[25] Martin Thoma. On-line Recognition of Handwritten Mathematical Symbols. PhD thesis, National Research Center, 2014.
[26] Edward Rolando Núñez Valdez, Juan Manuel Cueva Lovelle, Óscar Sanjuán Martínez, Carlos E Montenegro Marín, and Guillermo Infante Hernandez. Social voting techniques: a comparison of the methods used for explicit feedback in recommendation systems. International Journal of Artificial intelligence, IJIMAI, 1(2), pp.79, 2011
[27] Magdalena Wolska and Ivana Kruijff-KorBayova. Analysis of mixed natural and symbolic input in mathematical dialogs. 42nd Meeting of the Association for Computational Linguistics, 2004.
[28] Richard Zanibbi and Dorothea Blostein. Recognition and retrieval of mathematical expressions. International Journal on Document Analysis and Recognition, 2012.
[29] ClausZinn. A computational framework for understanding mathematical discourse. IGPL, 2003.
[30] L. de-la Fuente-Valentín, D. Burgos, R. González Crespo. A4Learning: Alumni Alike Activity Analytics to Self-Assess Personal Progress. IEEE International Conference on Advanced Learning Technologies, 2014

Alberto Corbi works as a senior researcher at the Technology-enhanced Learning & Social Networks (TELSOCK) research group and as a teaching assistant at the School of Engineering, both part of the International University of La Rioja. With a background in physics (M.Sc. in ocean-atmosphere interaction), computing and education, he is currently involved in research fields around recommender systems, eLearning standards and systems interoperability. Simultaneously, he is working on his PhD thesis around medical imaging at the Spanish Council for Scientific Research (CSIC).

Daniel Burgos works as Vice-Chancellor for Research & Technology and UNESCO Chair on eLearning at the International University of La Rioja (unir.net, research.unir.net). Previously he worked as Director of Education Sector and Head of User Experience Lab in the Research & Innovation Department of Atos, Spain, and as an Assistant Professor at Open Universiteit Nederland before that. His research interests are mainly focused on Adaptive and Informal eLearning, Learning & Social Networks, eGames, and eLearning Specifications. He is or has been involved in a number of European-funded R&D projects, such as Intuitel, Hotel, Edumotion, Inspiring Science Education Stellar, Gala, IntelLEO, Go-MyLife, Grapple, Unfold, ProLearn, TenCompetence, EU4ALL, NiHao, Kaleidoscope, Sister, and ComeIn. Prof. Dr. Burgos holds degrees in Communication (PhD), Computer Science (Dr. Ing), Education (PhD), and Business Administration. Furthermore, he is a member of a number of Executive Boards of associations and professional clusters focused on educational technology and eLearning innovation, for example Telearc, Telspain, Menon Network, Adie, and others.

Variation of the Heartbeat and Activity as an Indicator of Drowsiness at the Wheel Using a Smartwatch

Sergio Ríos-Aguilar , José Luis Miguel Merino, Andrés Millán Sánchez and Álvaro Sánchez Valdivieso

Escuela Superior de Ingeniería Informática y Arquitectura, Universidad Pontificia de Salamanca Madrid Campus, Spain

***Abstract* — Sleepiness is one of the first causal factors of accidents. An estimated 10-30% of road deaths are related to fatigue driving. A large number of research studies have been conducted to reduce the risk of accidents while driving. Many of these studies are based on the detection of biological signals by drowsiness/sleepiness. The activity of the autonomic nervous system (ANS) presented alterations during different physical states such as stress or sleepiness. This activity is measured by ECG (electroencephalogram) and, in different studies, it can be measured with the variation of the heart beat (HRV-Heart Rate Variability) in order to analyze it and then detect drowsiness/sleepiness in drivers. The main advantage is that HRV can be performed using non invasive and uncomfortable means compared to EEG sensors. New Wearables technologies are capable of measuring the heart beat and, further, using other sensors like Accelerometer and Gyroscope, embedded on a simple clock allow us to monitor the physical activity of the user. Our main goal is to use the pulsations measurements in conjunction with the physical activity for the detection of driver drowsiness/sleepiness in advance in order to prevent accidents derived from fatigue.**

***Keywords* — Drowsiness, sleepiness, smartwatch, sensors, heart rate, wearables**

I. INTRODUCTION

HISTORICALLY, the role of sleepiness and sleep disorders have been underestimated in comparison with other classic causes of accidents, such as alcohol or drugs, which may also be associated with drowsiness/sleepiness.

Previous studies estimate that between 10-30% of traffic accidents are directly related to sleepiness and fatigue while driving [1][2][3][4][5]. The fatigue related accidents are directly dependent on the time of day (circadian factor) and the type of road, particularly on motorways and monotonous roads. Mortality associated with such accidents is very high, probably because of the high speed at impact both the type of road and the lack of reaction by the driver.

Drivers usually are unable to assess their symptoms to detect or predict events drowsiness within minutes. This indicates that drivers may need technological interventions for detecting drowsiness/sleepiness while driving. Given these facts, we decided to find a method for alerting the driver about drowsiness/sleepiness. We had to find a way to evaluate driver fatigue states compatible with driving.

Studies related to ANS activity, detection of different physical states and physical activity detection were reviewed and we found that the current technology allowed to reconcile driving with a detection and warning of drowsiness/sleepiness mechanism.

This paper is structured as follows. After discussing related works in section 2, we describe in section 3 the methodology for this case study, technical details about extraction for classifying human activities. Finally section 4 presents the final conclussions

II. RELATED WORK

Several studies have been conducted to identify effective ways to monitor sleepiness driving [6]. The eye-blinking indicator has been the favorite in many studies. Almost in all the cases, the most commonly used test is the EEG. However, some uncomfortable EEG electrodes were used on the head, which are inappropriate for detecting drowsiness/sleepiness while driving, which is our goal.

Another biological measure that has been significantly related to drowsiness/sleepiness is Heart Rate Variability. HRV is regulated by both the sympathetic and parasympathetic systems, and it can be used to estimate the autonomous nervous system (ANS) activity [1][7]. The activity of the autonomous nervous system also can be measured in a non- invasive manner from the HRV, whose signal can be obtain from an ECG (electrocardiogram) [1].

DOI: 10.9781/ijimai.2015.3313

HRV has been studied in transitions from wakefulness to states of extreme relaxation, and there has been a decrease in Heart rate in the beginning of the dream [1][19].

Once it is known that the number of pulses decreases during sleep [8], the degree in which the number of pulses decrease during the non-REM stages of sleep has been explained as a depression in the activity of the sympathetic central system. Spectral analysis techniques have been used to link the pulse changes with the activity of the central nervous system [8].

Furthermore, there have also been studies in which are shown, it can be detected different patterns of physical activity [9-15][17][18] using the accelerometer and gyroscope sensors, one being the state of drowsiness/sleepiness. These studies were worth to know that it is possible to identify activities through sensors like the accelerometer and the gyroscope. Among them are various studies attempting to interpret body activity using the latest current technology (wearables) [9]. More specifically used smart watches, for example the Samsung Gear S, using accelerometer and gyroscope sensors [10], and others also include pedometer [11][20][21][23].

All these studies clarify the possibility of detecting sleep or detect patterns of physical activity. Our study goes a step further and has combined both formulae.

III. METHODOLOGY

Assuming you could use data obtained from Heart Rate Monitor (HRM) to detect changes in the nervous system and on the other hand, the data obtained from accelerometer, gyroscope and pedometer to identify physical activities, we decided to unify all the data in order to detect the possible state of drowsiness/sleepiness of a driver, and then alert the user advising him to stop driving.

For this experiment, we needed to find a device that contains all the necessary sensors and did not pose a threat to the user when driving. Technology currently offers, in a single device, a smartwatch: information provided by the heart rate and also accelerometer, gyroscope and pedometer sensors to detect as accurately as possible states of drowsiness/sleepiness during driving. In a smartwatch, the heart rate sensor is PPG type (photoplethysmography). These sensors use a light-based technology to sense the rate of blood flow as controlled by the heart's pumping action. The other sensors are similar to other devices like tablets or smart phones.

For our choice, the essential features that should be fulfilled by the smart watch include those aforementioned sensors and large battery capacity, since the continued use of all the sensors causes a non insignificant power consumption. Other features, such as screen size, were not significant for the election. Several Smart watches, like Sony Smart Watch 3 SWR50, LG G Watch o Motorola Moto 360, were discarded because they do not have HRM or other sensors.

The smart watch chosen was the Samsung Gear S, containing all the above mentioned sensors, battery life of 4 days with occasional use, and 1-2 days with increased use, and allows us to program the required application to use all the information.

It should be noted that the study of drowsiness/sleepiness at the wheel requires a deep and thorough knowledge of the behavior of people. It is therefore necessary for the study to collect data from both sexes and different ages, ranging from 18 to 70 years, since the behavior with respect to sleep and, especially bodily activity, is different in each of them and reaches the state of drowsiness slightly differently. Another fact to note are the types of roads; the motorways require a monotonous driving as indicated in studies [2] y [5] it induces sleep and causes more accidents, and this is considered in this study.

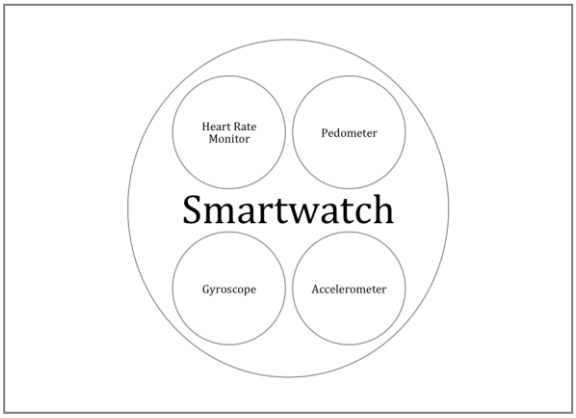


Fig 1. Smart watch sensors.

On the other hand, the fact of using a smart watch (Fig 1)[11] that is worn on the wrist and contains sensors Herat rate, accelerometer, gyroscope and pedometer, has many advantages as all sensors are integrated in one device, it is easy to carry, and also allows to detect the lack of physical activity while driving. But, it makes it so challenging the classification of physical activity since the smart watch worn on the wrist should control very well the movements made by hand to predict the driver's state [13][16].

A. Data Acquisition/collection.

In this first phase, we gathered all the information that sensors reported to us from the largest possible number of people, with different ages and in different scenes in order to see the behavior and so we can analyze data and classify the states of people.

We invited 30 volunteers to collected data, with equal numbers of men and women, ages commented above. The Gear S smart watches were prepared to store the user data for the following case studies:

- Driving his/her usual vehicle under normal conditions of their day to day. Samples of city driving.
- Driving his/her vehicle at motorway, minimum distances of 30 minutes. Samples of short trips on the highway.
- Driving his/her usual car on long trips, doing break every two hours. Samples of long trips on the highway.

- Being tired, driving in a driving simulator (City Car Driving + Steering Wheel Logitech G27) for two hours, simulated driving situation on a highway.
- Watching TV before sleeping.
- Going to bed and sleeping the whole night.

The first three scenarios are real driving. The other scenarios may not be exactly like a real driving situation, but the use of simulation is pretty close to reality and, they serve to not risk the integrity of individuals and to gather some subjective outcomes.

After completing driving, the driver filled out a form with the actions that had performed, the route taken and, if there was anything significant related to their physical state (either, tired, fatigued, some drowsiness, stressed, etc). Thus, we obtained daily data about the behavior of different people, as well as, samples that we knew what behavior was followed by the driver and under what conditions as well.To detect cases of sleep, the person took note of the time that he slept and the time he started to watch the television.

The objective of this first phase is the collection of samples that allow us to analyze the data for the unambiguous detection of the different states.

The data collected are comprised:

- Values of the 3-axis accelerometer (20 values / sec).
- Values of the 3-axis gyroscope (20 values / sec).
- Values of the pedometer every second.
- Values of HRM every second (Beats and peak to peak).

B. Feature Extraction.

For each of the samples obtained, we analyzed on the one hand the results collected by the accelerometer, gyroscope and pedometer sensors; on the other hand, the results collected by the heart rate sensor [22],[23].

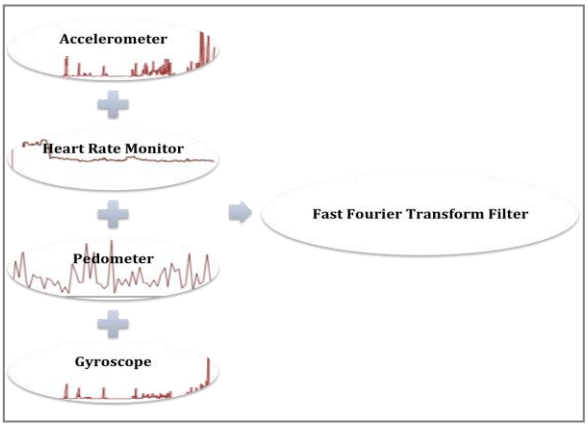


Fig 2. Signal analysis

Using the appropriate tools (pic 2, FFT-Fast Fourier Transform) and details of actions such as traveling and fitness reported by the driver, we analyzed the data to identify the driver´s actions, that are:

- Walking to the car.
- Getting in the car.
- Leaving the parking.
- Driving.
- Parking.
- Getting out of the car.

After analyzing the results of each driver, they analyzed as a whole, and common patterns were found in the results captured for the different states of the driver. There were small differences between each of them, but it could be possible to clearly identify the different actions.

C. Classification.

Using FFT tools, various types of filters and scatter plots points, we found out common patterns for each of the states of the driver and state of drowsiness/sleepiness. While driving behavior of each person is not the same, after analyzing the data, with very small differences, they all converge at the same point for each of the states identified [10] [12]. We can find more nervous drivers when parking -since they made faster movements-, yet, other drivers made movements with greater confidence, but in the end, both drivers make the same set of movements (82%) and this is what serves our interests. In the case of the detection of sleep, the same approach is followed. Although there are people who find it hard to get more sleep and other costs them less, ultimately all of their heart rate converge to decrease rates in order to stabilize at a threshold associated with the dream state. For those cases, where the person fell asleep, such as going to bed or get to watch TV after dinner, it observed as pulsations are reduced as was reconciling the dream, as other studies have reported [1][19]. It is also noted that the values set with the accelerometer and gyroscope are few swings because the driver tends to correct the direction of the car compared to previous moments.

The data were separated and classified according to the following states:

- Walking.
- Steady state.
- Excitation.
- Driving.
- Parking maneuver.
- Sleepiness.
- Slumber.
- Deep sleep.

D. Process.

After analysis of the data with the various tools, such as Kalman filter, pass band filter, high pass filter…, we concluded that these data did not provide further information to identify actions, which were ruled to implement the system.

The software system developed for detecting driver drowsiness/sleepiness, and providing additional capabilities, it´s based on the following processing:

To identify actions using data collected by the sensors accelerometer, gyroscope y pedometer:

Fourier series:

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right)$$

Fast Fourier Transform (FFT):

$$f_i = \sum_{k=0}^{n-1} x_k \, e^{-\frac{2\pi i}{n} jk} \qquad j = 0, \ldots, n-1$$

Vector direction:

$$\left| \overrightarrow{PQ} \right| = \sqrt{\overrightarrow{x^2} + \overrightarrow{y^2} + \overrightarrow{z^2}}$$

For identification of drowsiness with heart rate sensor data:
Average:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + a_3 + \ldots + a_n}{n}$$

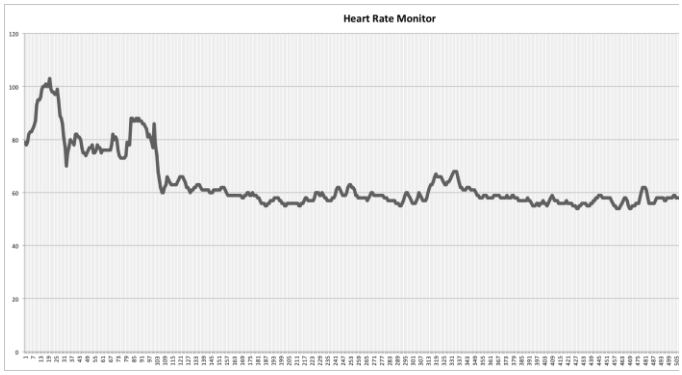


Fig 3. Heart Rate Monitor

The algorithm processes the data stored once it has completed the minimum time window for collecting the data set needed to predict the status of the driver. The window time is 30 seconds. At the beginning of the application, within 30 seconds, the algorithm is not run as the data set needed to decide the status of the driver is still not available. After the first 30 seconds, the algorithm, every second, runs checking the new data that have been registered (20 new records per second).

The final decision for detection of sleep is made based on the result (pic 4) by the algorithm which uses the heartrate sensor (50%) and the algorithm that uses accelerometer and gyroscope sensors (50%). Both must have had the same result, that is to say, having detected that the driver is entering a state of drowsiness/sleepiness, to alert the driver. Thus, false positives are avoided.

In addition, data collected from the accelerometer and gyroscope sensors are also used to detect other driver actions: park, leaving the car, etc. In the tests carried out, it should be noted, that in travel four hours, the battery has an average consumption of 50%, running throughout the trip application in background mode.

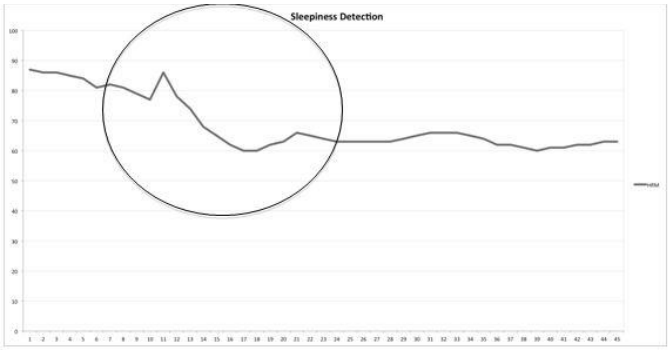


Fig 4. Sleepiness detection

E. Experimentation.

Several methods are designed to experience algorithms to detect sleep:

1. Online method: once the algorithm was coded, the resulting application is installed on the smart watch and re-run on the same users on which data were captured. The application runs in debug mode so that we could record all the events produced and re-analyze the data recorded to refine, if necessary, the algorithm.
2. Offline method: Having already recorded data from previous lines, what we did was to define a test environment using the same algorithm we have designed for the application, so you can run the same data stream on the algorithm in real time and then you can also validate the algorithm designed. Thus, also we can detect errors of the algorithm in a fully controlled environment that simulates the real environment.

IV. Conclusion

In this paper, we have presented a new application designed for Samsung Gear S smart watch recognizing and alerting drowsy state of a driver at the wheel of a vehicle. Using this application is to avoid car accidents caused by the driver falling asleep at the wheel.

Other application functionalities are (1) warning to the driver that he must rest every two hours, and (2) firing an alarm in states of nervousness that could mean danger while driving.

In this application, specific algorithm was designed using known mathematical tools such as FFT, filters, scatter of points and statistic power in order to detect that the driver is entering a state of drowsiness/sleepiness. It has a very acceptable power consumption rate and at no time blocks another applications or the system itself, thus providing optimum performance.

Finally. using this application on Samsung Gear S smartwatch is a very simple and easy way, from user point of view, to be protected against a circumstance of drowsiness/sleepiness while driving.

REFERENCES

[1] Vicente, J., Laguna, P., Bartra, A., & Bailon, R. (2011). Detection of driver's drowsiness by means of HRV analysis.2011 Computing in Cardiology (CinC), 89–92.

[2] Horne, J. a, & Reyner, L. a. (1995). Sleep related vehicle accidents. BMJ (Clinical Research Ed.), 310, 565–567. doi:10.1136/bmj.310.6991.1411b

[3] Epidemiological, A., & Study, M. (1998). Automobile Accidents in Patients with Sleep Apnea Syndrome.

[4] Leger, D. (1994). The cost of sleep-related accidents: a report for the National Commission on Sleep Disorders Research. Sleep, 17(2), 84–93.

[5] Horne, J., & Reyner, L. (1999). Vehicle accidents related to sleep: a review. Occupational and Environmental Medicine, 56, 289–294. doi:10.1136/oem.56.5.289

[6] Mahachandra, M., Yassierli, Sutalaksana, I. Z., & Suryadi, K. (2012). Sensitivity of heart rate variability as indicator of driver sleepiness.2012 Southeast Asian Network of Ergonomics Societies Conference: Ergonomics Innovations Leveraging User Experience and Sustainability, SEANES 2012, 0–5. doi:10.1109/SEANES.2012.6299577

[7] Elsenbruch, S., Harnish, M. J., & Orr, W. C. (1999). Heart rate variability during waking and sleep in healthy males and females. Sleep, 22(8), 1067–1071.

[8] Bonnet, M. H., & Arand, D. L. (1997). Heart rate variability: Sleep stage, time of night, and arousal influences. Electroencephalography and Clinical Neurophysiology, 102(96), 390–396. doi:10.1016/S0921-884X(96)96070-1

[9] Ayu, M. A., Ismail, S. A., Matin, A. F. A., & Mantoro, T. (2012). A Comparison Study of Classifier Algorithms for Mobile-phone's Accelerometer Based Activity Recognition. Procedia Engineering, 41(Iris), 224–229. doi:10.1016/j.proeng.2012.07.166

[10] Mortazavi, B. J., Pourhomayoun, M., Alsheikh, G., Alshurafa, N., Lee, S. I., & Sarrafzadeh, M. (2014). Determining the single best axis for exercise repetition recognition and counting on smart watches. Proceedings - 11th International Conference on Wearable and Implantable Body Sensor Networks, BSN 2014, 33–38. doi:10.1109/BSN.2014.21

[11] Bieber, G., Haescher, M., & Vahl, M. (2013).Sensor requirements for activity recognition on smart watches.Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments - PETRA '13, 1–6. doi:10.1145/2504335.2504407

[12] Brezmes, T., Gorricho, J., & Cotrina, J. (2009). Activity Recognition from Accelerometer Data on a Mobile Phone.Test, 5518, 796–799. doi:10.1007/978-3-642-02481-8_120

[13] Lau, S. L., & David, K. (2010). Movement recognition using the accelerometer in smartphones. 2010 Future Network & Mobile Summit, 1–9. doi:http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5722356

[14] Casale, P., Pujol, O., & Radeva, P. (2011). Human activity recognition from accelerometer data using a wearable device.Pattern Recognition and Image Analysis, 289–296. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-21257-4_36

[15] Maguire, D., & Frisby, R. (2009). Comparison of Feature Classification Algorithms for Activity Recognition Based on Accelerometer and Heart Rate Data. 9th. IT & T Conference.

[16] Yang, J. Y., Wang, J. S., & Chen, Y. P. (2008). Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers. Pattern Recognition Letters, 29(16), 2213–2220. doi:10.1016/j.patrec.2008.08.002

[17] Fuentes, D., Gonzalez-Abril, L., Angulo, C., & Ortega, J. a. (2012). Online motion recognition using an accelerometer in a mobile device.Expert Systems with Applications, 39(3), 2461–2465. doi:10.1016/j.eswa.2011.08.098

[18] Fahim, M., Fatima, I., Lee, S., & Park, Y. T. (2013). EFM: Evolutionary fuzzy model for dynamic activities recognition using a smartphone accelerometer. Applied Intelligence, 39(3), 475–488. doi:10.1007/s10489-013-0427-7

[19] Hayata, Y., Bhuiyan, S., Kawanaka, H., & Oguri, K. (2013). Prediction of the Time When a Driver Reaches Critical Drowsiness Level Based on Driver Monitoring Before and While Driving, (Itsc), 6–11.

[20] Fernández MJL, Fernández JG, Rios-Aguilar S, BS Selvi, RG Crespo Control of attendance applied in higher education through mobile NFC technologies. Expert Systems with Applications 40 (11), 4478-4489

[21] JP Espada, VG Díaz, RG Crespo, OS Martínez, BCP G-Bustelo Using extended web technologies to develop Bluetooth multi-platform mobile applications for interact with smart things. Information Fusion 21, 30-41

[22] Rios-Aguilar S., R González-Crespo, L de-la-Fuente-Valentin. Anonymous Assessment Information System for Higher Education Using Mobile Devices Advanced Learning Technologies (ICALT), 2014 IEEE 14th International.

[23] JP Espada, V Garcia Diaz, R Gonzalez Crespo, PG Bustelo, B Cristina. An intelligent Mobile Web Browser to adapt the mobile web as a function of the physical environment.Latin America Transactions, IEEE (Revista IEEE America Latina) 13 (2), 503-509

Sergio Ríos Aguilar holds a PhD in Economics from University of Granada. He also holds a MSc in Telecommunications from King Juan Carlos University (URJC) and Computer Sciences degree from UOC. He currently works at UPSAM teaching Mobile Development and is conducting research on advanced mobile enterprise information systems.

Jose Luis Miguel Merino holds a degree in Systems Engineering by the UNED, and works as an IT Consultant at Pavoni School and undertakes freelance IT projects.Computer Sciences graduate by the Pontificial Universty of Salamanca, Madrid Campus. School of Engineering.

Álvaro Sánchez Valdivieso received his B.Sc in Computer Science (2001) at Universidad Pontificia de Salamanca (Madrid). Recently, finished his academic degree at Universidad Pontificia de Salamanca as well. Working on simulation software projects for aeronautical industry since 2002, latest years, his researches are focused on health applications for wearables and smart phones. Skilled on device sensors.

Andrés Millán Sánchez was born in Madrid in 1973. He studied Engineering in Computer Systems at the "Universidad Pontificia de Salamanca en Madrid". Since 1998 works as Head of Computing Department in an automotive company in Madrid (Movilnorte, S.A.).

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?

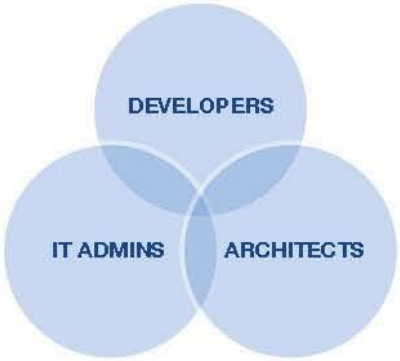
ElasticBox Solution

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