Applications of Language & Knowledge Engineering

Research in Computing Science

Series Editorial Board

Research in Computing Science es una publicación trimestral, de circulación internacional, editada por el Centro de Investigación en Computación del IPN, para dar a conocer los avances de investigación científica y desarrollo tecnológico de la comunidad científica internacional. **Volumen 145,** noviembre 2017. Tiraje: 500 ejemplares. *Certificado de Reserva de Derechos al Uso Exclusivo del Título* No.: 04-2005-121611550100-102, expedido por el Instituto Nacional de Derecho de Autor. *Certificado de Licitud de Título* No. 12897, *Certificado de licitud de Contenido* No. 10470, expedidos por la Comisión Calificadora de Publicaciones y Revistas Ilustradas. El contenido de los artículos es responsabilidad exclusiva de sus respectivos autores. Queda prohibida la reproducción total o parcial, por cualquier medio, sin el permiso expreso del editor, excepto para uso personal o de estudio haciendo cita explícita en la primera página de cada documento. Impreso en la Ciudad de México, en los Talleres Gráficos del IPN – Dirección de Publicaciones, Tres Guerras 27, Centro Histórico, México, D.F. Distribuida por el Centro de Investigación en Computación, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, México, D.F. Tel. 57 29 60 00, ext. 56571.

Editor responsable: *Grigori Sidorov, RFC SIGR651028L69*

Research in Computing Science is published by the Center for Computing Research of IPN. **Volume 145** November 2017. Printing 500. The authors are responsible for the contents of their articles. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of Centre for Computing Research. Printed in Mexico City, in the IPN Graphic Workshop – Publication Office.

Applications of Language & Knowledge Engineering

Beatriz Beltrán
Darnes Vilariño
Josefa Somodevilla
David Pinto

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2017

ISSN: 1870-4069

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx

Indexed in LATINDEX, DBLP and Periodica

Printing: 500

Printed in Mexico

ISSN 1870-4069

Editorial

This volume of the "Research in Computing Science" journal contains selected papers related to the topic of Language and Knowledge Engineering and their applications. The papers were carefully chosen by the editorial board on the basis of the at least two double blind reviews by the members of the reviewing committee or additional reviewers. The reviewers took into account the originality, scientific contribution to the field, soundness and technical quality of the papers. It is worth noting that various papers for this special issue were rejected (rejected rate was 25%).

The volume contains a selection of papers presented in the 5th International Symposium on Language & Knowledge Engineering (LKE'2017), an academic conference organized in the Faculty of Computer Science of the Benemérita Universidad Autónoma de Puebla (BUAP) which has been created and organized for the fourth time by the Language & Knowledge Engineering Lab with the aim of offering an academic platform in which experts in related areas may exchange experiences and publish their recent research advances.

We would like to thank Mexican Society for Artificial Intelligence (Sociedad Mexicana de Inteligencia Artificial) and the Thematic Academic Network named "Language Technologies" (Red Temática en Tecnologías del Lenguaje) for their invaluable support in the construction of this volume.

The entire submission, reviewing, and selection process, as well as preparation of the proceedings, were supported for free by the EasyChair system (www.easychair.org).

Beatriz Beltrán
Darnes Vilariño
Josefa Somodevilla
David Pinto
Guest Editors
Benemérita Universidad Autónoma de Puebla,
LKE-FCC-BUAP, Mexico

November 2017

ISSN 1870-4069

Table of Contents

Page

ISSN 1870-4069

Medical System for X-Ray Analysis

Bárbara Emma Sánchez[1], Alberto Jaramillo[2], Guillermo Vara[1]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de Computación, Puebla, México

[2] INAOE, Luis Enrique Erro Tonanzintla, Puebla, México

brinza@hotmail.com

Abstract. Problems within the bone system that a patient experiences when undergoing cancer treatment are always a cause for concern, because they can trigger other serious diseases such as osteoporosis , among others, due to the loss of bone density in the system, caused by bone metastasis, that is why it is necessary to give them a detailed follow-up and apply different treatments to avoid this type of problems, for this reason the main reason for the development of a system of this kind that allows to have better efficiency at the time of medication and both the patient and the specialist doctor are more certain that the results are obtained from a system that allows to calculate the difference between two images, one taken at the beginning of the treatment and a later time, obtaining graphs And numerical percentages that the medical specialist can understand to make a better decision knowing that's on from a reliable and secure source.

Keywords. System, rotations, histograms, radiography, Python.

1 Introduction

The present project develops a visual medical system for the optimization of the medication process to patients with cancer problems, who have been prescribed a treatment to avoid the deterioration of the bones caused by the aforementioned disease.

Develop a medical system under the Python programming language that works as an aid to doctors in determining if a treatment is working on a patient who is undergoing certain treatment.

Has as purpose:

- Time optimization for doctors and users.
- Assist in the diagnosis of medical treatments.
- Improve the care and efficiency of diagnosis.

For the programming of the medical system in its version 1.0, the system was developed on the programming language Python which is a language that has efficient data structures, high level and a simple but effective approach in object programming which

makes it a language suitable for the realization of the system and due to the great support with which it has within the community has an extensive number of libraries to support when processing and interpreting the images.

The medical system was developed on a computer with the following characteristics:

- Operating System: Windows 10 x64 with 8 GB Ram,
- Python: 2.7.13,
- OpenCV: 3.2,
- PyInstaller: 3.2,
- Numpy: 1.3,
- IDE: SublimeText 3126 for 64-bit.

Below is a brief description of the elements and libraries that were necessary to install to carry out the programming.

2 Starting the Program

From the wx.Python library for window management, we create the startup window where the user is presented with the system (Fig. 1) and invited to enter [1].

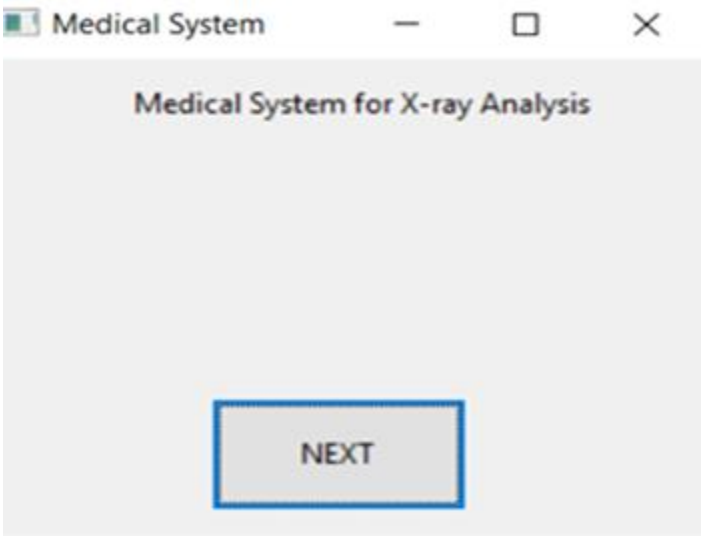

Fig. 1. Start-up interface.

Once the user has entered the system, a user interface (Fig. 2) has been created that has different options for the use of the software. In order to begin to occupy the software correctly, it is always necessary to have the images loaded otherwise the system will throw an error.

Where the buttons are created for operations from the wx.Buttons classes and are added to the interface with the "self.sizer.Add ()" method that is obtained from the main

interface named "self" and the class "Sizer" from which the "Add" method is retrieved, which receives the widget information that will be added to the main window, as shown in figure 2.

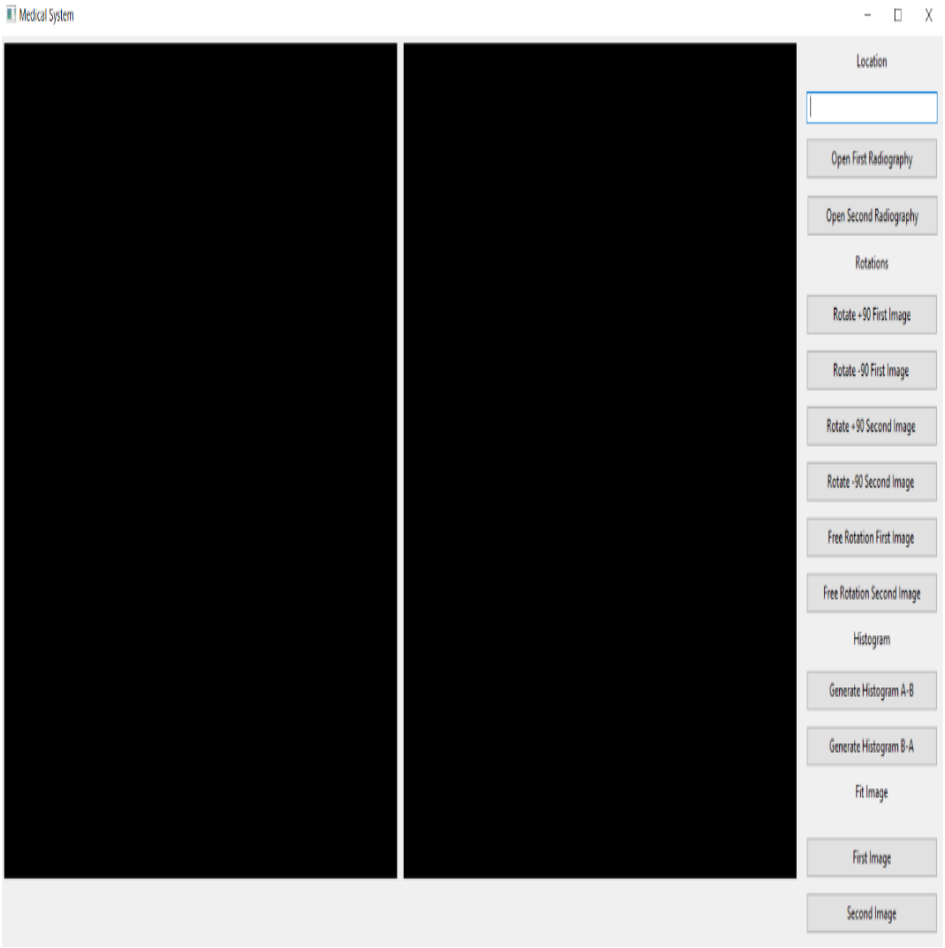

Fig. 2. Main interface.

Then with wx.Python we create from the left side what are the windows of 600x600 where the x-rays chosen by the user in a * .jpg format are uploaded, these images can be loaded with the buttons that are on the right side where they are Indicates to which image box is destined, to be able to load the image the OpenCV library was used to be able to pass the image of a format * .jpg to a format BitMap with which you can carry out operations later, Image to the size of the aforementioned box, in order to make this adjustment and not lose the proportions of the image.

This allows the user to present an image adapted to 600x600 without losing the appearance of the image (Fig. 3).

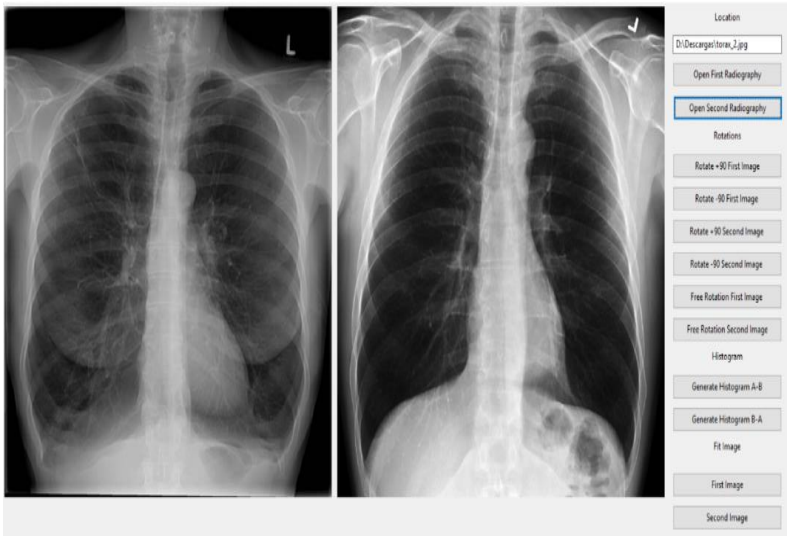

Fig. 3. Presentation of the X-ray.

3 Operations on the Images

3.1 Rotations

Within the system interface and for greater convenience when working with x-rays, the system allows rotations of 90 °, -90 ° and 0-360 ° (Fig. 4), so that at the moment to load two images and need to adjust the position of some of them for their correct analysis can be performed without having to occupy any external tool. [2].

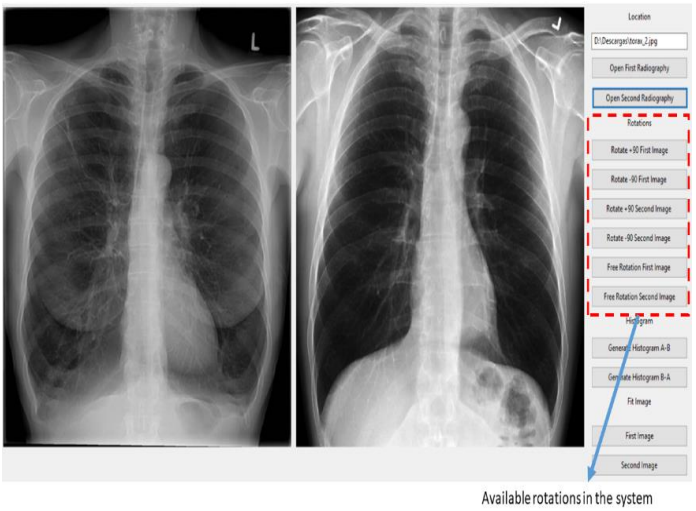

Fig. 4. Rotations available.

To rotate 90 and -90 ° to use these functions, it is only necessary to press one of the buttons "Rotate 90" (Fig. 5) depending on the case and the image to be rotated, it must be taken into account that once. That the rotation operation is performed on the image is over writes the original image for use in the next operation, so it is advisable to make a backup of your original file before performing any operation with the image.

The implementation of the aforementioned rotations, were performed through OpenCV by means of the function "getRotationMatrix2D (center, angle, scale)", which asks us for two points corresponding to the center of the image which were calculated and saved in a tuple by calculating them from the division of the width of the image between two for the first point and the height of the image for the second point.

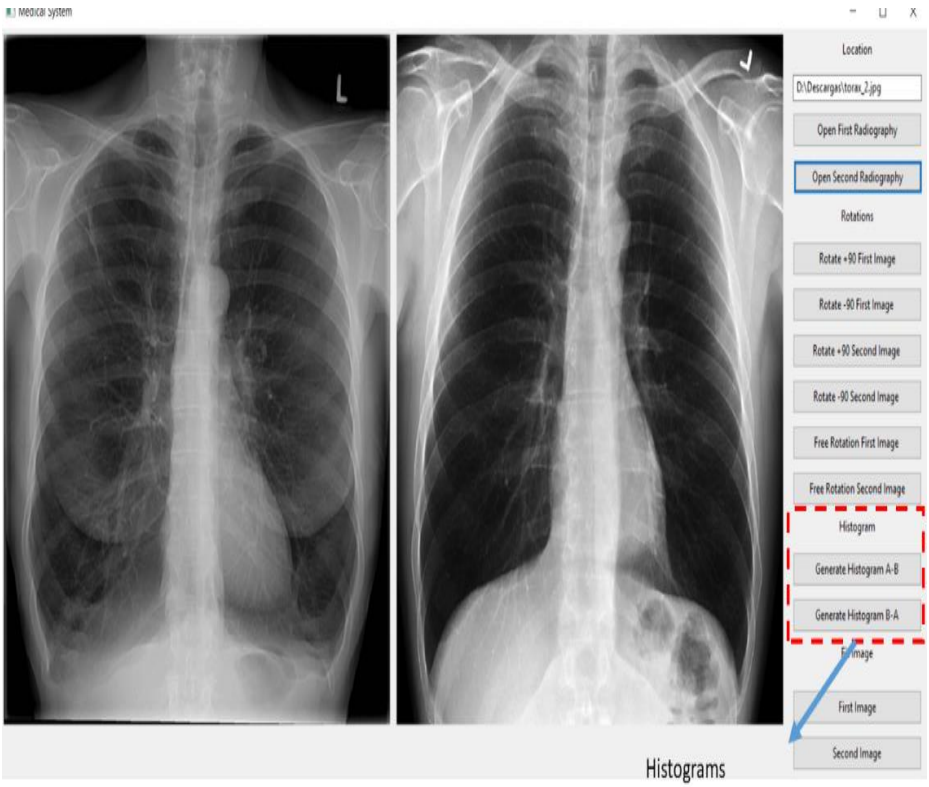

Fig. 5. Default Rotations.

To use the free rotation functions, it is only necessary to press one of the buttons "Free rotation" (Fig 6) depending on the case and the image to be rotated, once the button is pressed the system launches (Fig. 7) where it is necessary to indicate by degrees the rotation in which the degrees of rotation go from 0° to 360° taking into account that all rotation goes in the direction of the clock [3] .

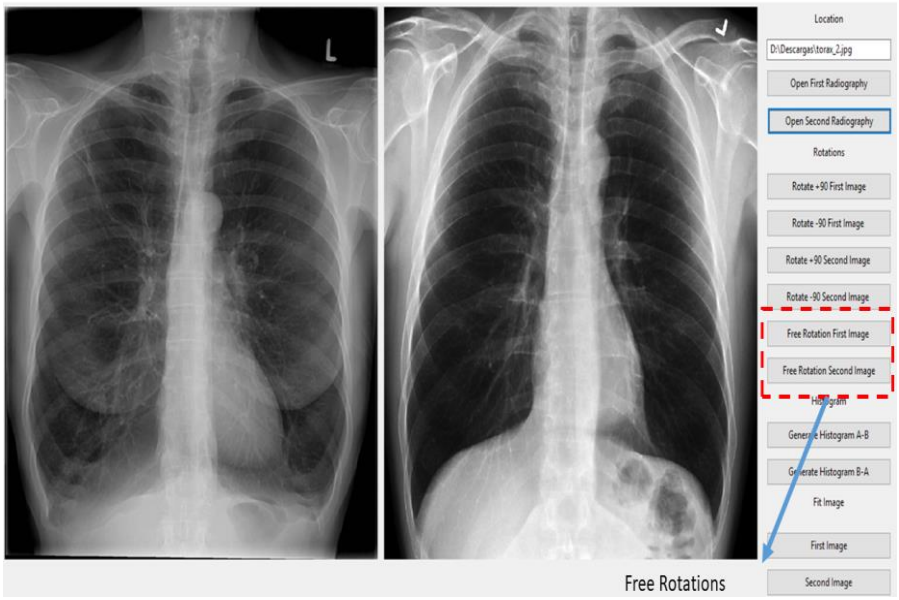

Fig. 6. Free rotation.

It is necessary to take into account that once the operation of rotation on the image is performed, the image is written to originate for using in the next operation, therefore it is advisable to make a backup of its original file before performing any operation with the image.

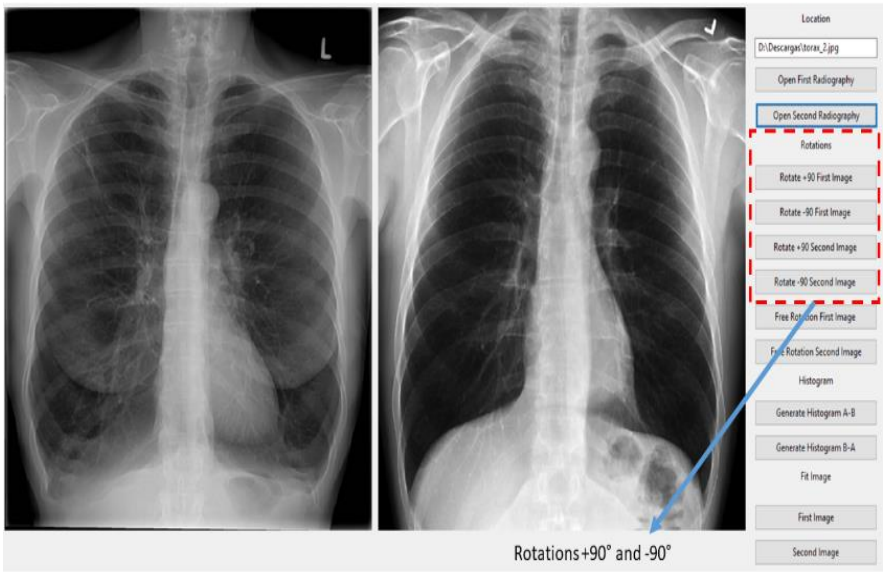

Fig. 7. Rotational angles.

For the implementation of the free rotation, at the time of launching the window to indicate the degrees of rotation, a new window was created with wx.Python which receives the values for the rotation and once the user clicks "Accept "The system makes a step of parameters towards the corresponding method which receives them and passes them to the function of OpenCV that by means of the function of" getRotationMatrix2D (center, angle, scale) ", performs the rotation.

4 Adjustment on the Imagen

The use of this tool was developed with the aim of allowing users to trim the image in a given case that the image has parts that will not be relevant when performing our analysis or in case you only want to analyze certain part of the image and discard some other. To use this tool, we have to go to the "Image Adjustment" section (Fig. 8) and select the image that we want to adjust for which we skip a window (Fig. 9) with the image to be worked once we have that window, we proceed to press the beginning of where we want to cut and drag without releasing the mouse towards the other end that we want to have of our image [1,2].

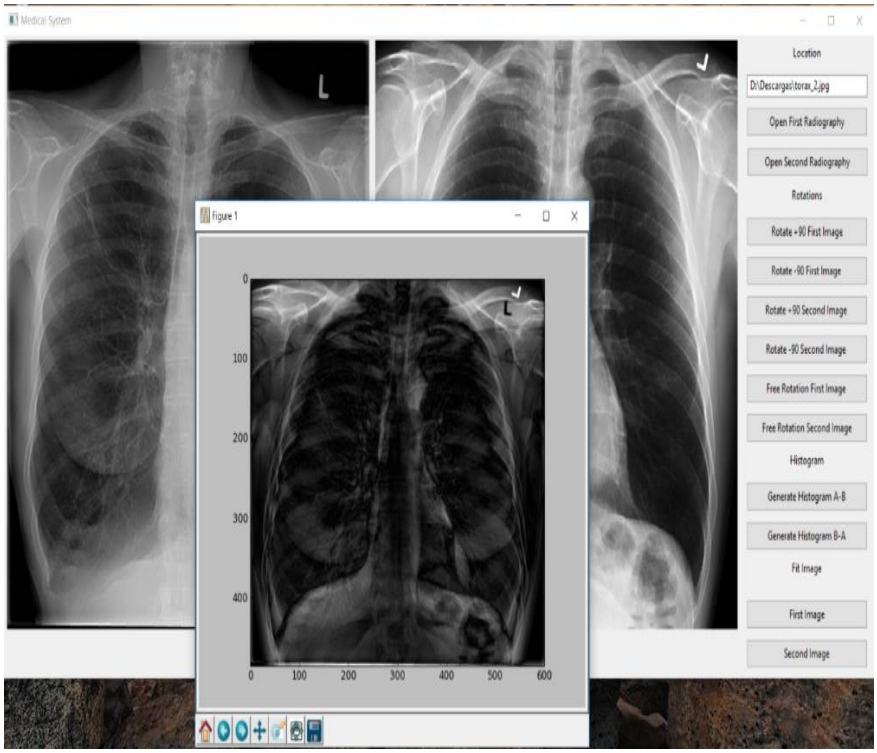

Fig. 8. Adjusting images.

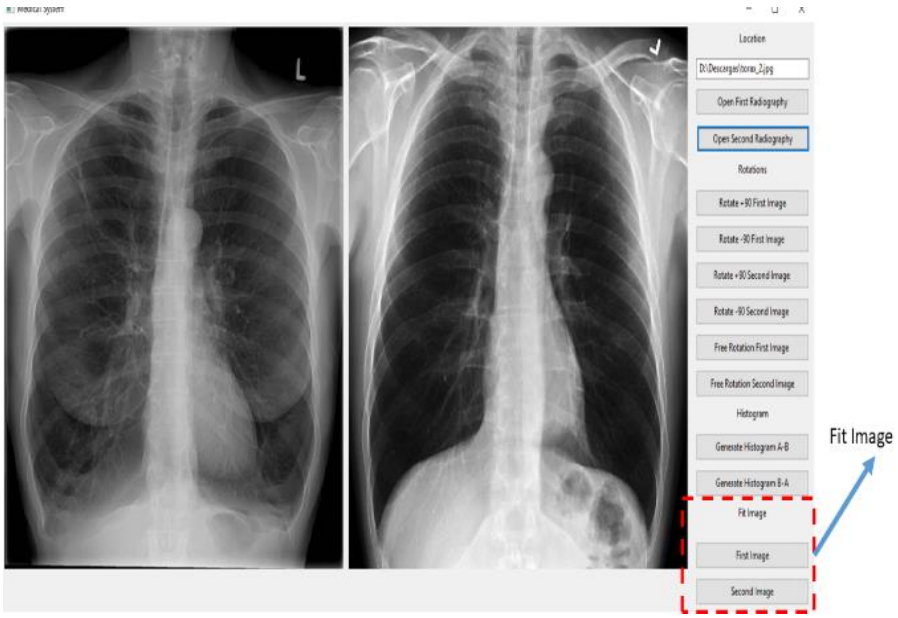

Fig. 9. Picture adjustment interface.

For the development of this section the OpenCV library was occupied, first what was done was to initialize a window in which the user through a pointer begins to select the point at which point he wants to cut. The image, once the user releases the right click of the mouse as an indication that all points were captured the system through the "setMouseCallback" method obtains the coordinates of the points and save them in an arrangement that we pass to the method " Get_clicks () "where from the previously obtained points a cut of the image of rectangular form is made and we processed the points with the method" resize () "that returns to us the image cuts that we save in a temporary image variable that then it is opened directly in the main interface of the system where from then on you can work with it.

Once we release the right mouse button the system will crop the image and send it to the main window of the system to continue working with it, taking into account that the initial image will be overwritten so it is recommended to have a copy of the main image by If at some point it is necessary to reuse.

5 Calculate Histograms and Bone Gain

Once you have the TWO images already worked, you can calculate the histograms depending on the needs we can calculate "AB" or "BA", to perform these operations we located in the "Histogram" section (Fig. 10) and press the operation to be calculated

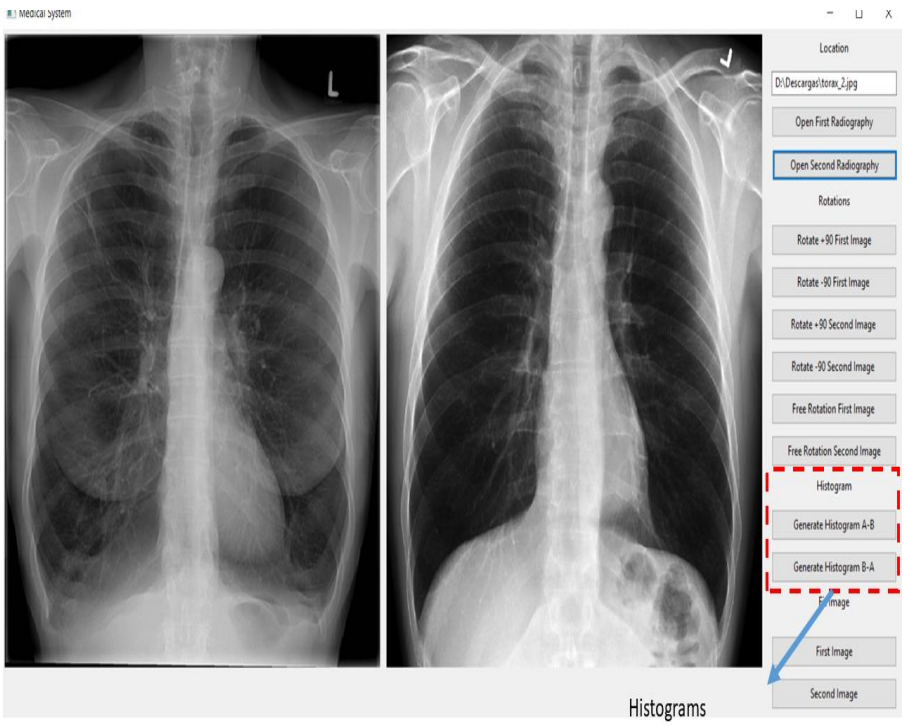

Fig. 10. Histograms.

The system calculates the histograms and the bone gain and it unfolds as follows taking into account that to calculate the histograms and the bone gain:

- Image subtraction result window (Fig. 11) in this window shows the resulting image of the subtraction and in this window you can perform operations such as "Save the image", "Configure the graphic", "Perform Zoom in a certain area "," Move the image within the graph "and" Return to its initial state ".

The implementation of the subtraction of the images was developed from the OpenCV library with the "absdiff" method, which is passed as parameters the two images that the system previously verified to be of the same size and That are in valid pixel formats. Once the subtraction of the images has been obtained, a "cv2" matrix is obtained which we convert into an image of type "wx.Image" so that it can be displayed inside the "plt" window which allows us to perform the Save, zoom and reset operations

- Results window of the bone gain (Fig. 12) this window will first show the bone gain in percentage, way followed by the average value of the gray tones of the resulting image and finally the average value of the gray tones of the initial image, taking as the initial image in case of "AB" the image "A" and in the case of "BA" the image "B". To calculate the bone gain, the matrix "cv2" is

traversed in the rows and columns to obtain the average gray scale for each one and to be able to apply the formula:

$$Bone\ gain = \frac{average\ value\ of\ the\ gray\ tones\ of\ the\ RTE\ image}{average\ value\ of\ the\ gray\ tones\ \ of\ the\ inicial\ image} * 100.$$

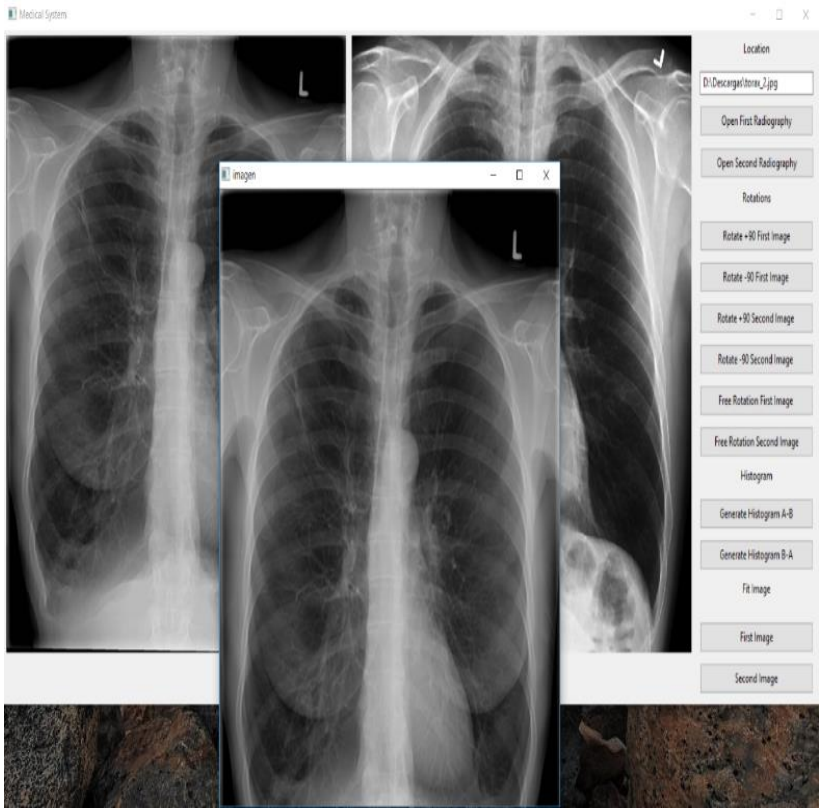

Fig. 11. Image substraction.

Histogram window (Fig. 13) this window will show the calculated histogram of the resulting image where the number of pixels is represented on the "Y" axis and the pixel's gray hue as the "X" axis. Within this window you can perform operations such as "Save the image", "Configure the graph", "Zoom into a specific area", "Move the image within the graph" and "Return to its initial state".

The histogram is calculated with OpenCV and the "calcHist" method, which is passed the image resulting from the subtraction and the range of values of the colors of the image, taking the case of the x-rays that are worked on scale gray is passed a range of 0-256 in grayscale.

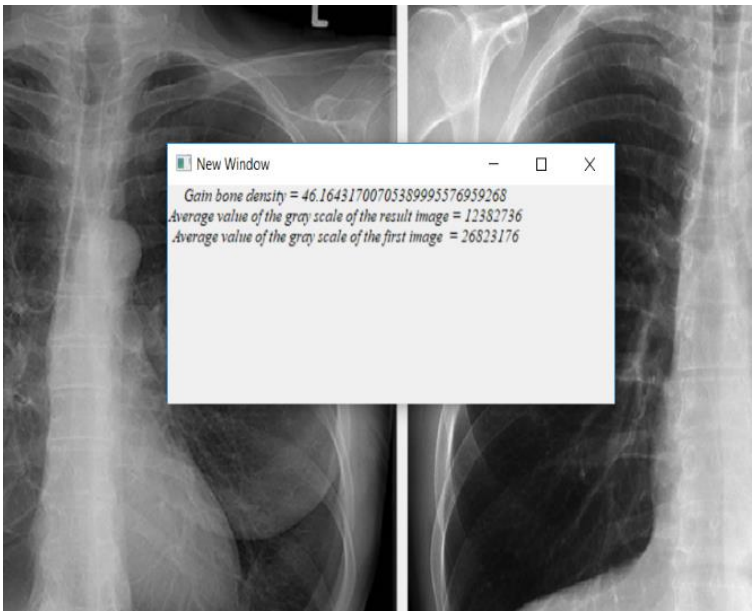

Fig. 12. Bone gain.

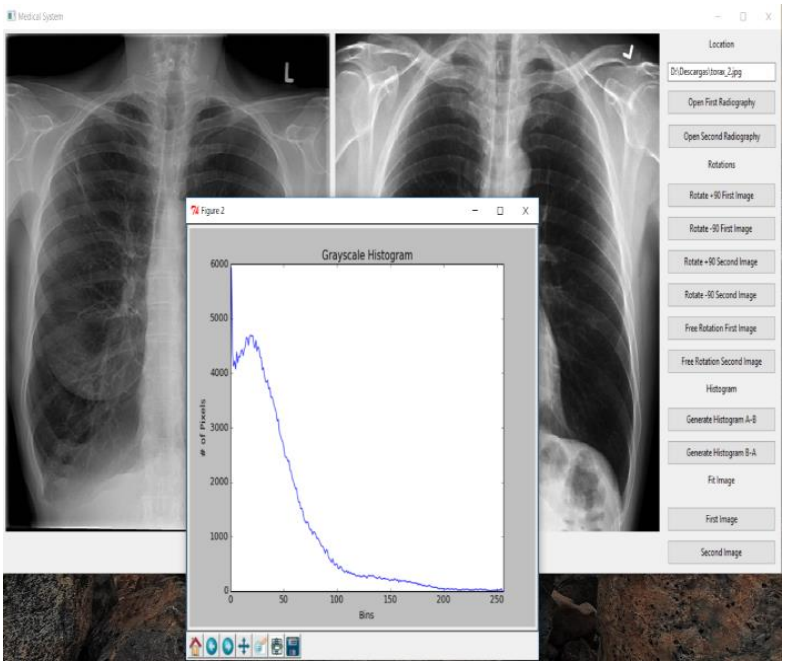

Fig. 13. Resulting image histogram.

6 Conclusions

So far the results obtained with the software have been satisfactory obtaining graphs and percentages that the doctors have been able to interpret, thus fulfilling the objective of the work, which is the optimization of time for both specialist doctors and patients. And making the method of drug delivery more efficient.

According to the results obtained, it is concluded that images taken by radiography can be processed by a computer to obtain more efficient numerical values that help the decision making by the doctors.

References

1 Reiser, M., Vahlensieck, M.: Resonancia Magnética Musculoesquelética. España: Médica Panamericana (2010)
2 Bohem, B.: Spiral Model of Software Development and Enhancement. IEEE Computer, May (1988)
3 https://www.blog.pythonlibrary.org/about/

ISSN 1870-4069

Partitioning Strategy Divide & Conquer as CPANs: A Methodological Proposal

Mario Rossainz[1], Manuel Capel[2], Diego Sarmiento[1]

[1]Benemérita Universidad Autónoma de Puebla,
Faculty of Computer Science, Mexico

[2]University of Granada, Software Engineering Department,
Granada, Spain

rossainz@cs.buap.mx,manuelcapel@ugr.es,diegorojas0888@gmail.com

Abstract. This work proposes the use of Parallel Objects using the High Level Parallel Compositions or CPAN model (Acronym in Spanish), to implement the communication patterns between processes most used in solving parallel problems. Particularly the implementation of the partitioning strategy divide and conquer as a CPAN using the object orientation paradigm is shown. A CPAN comes from the composition of a set three object types: An object manager that controls a set of objects references, which address the object Collector and several Stage objects and represent the CPAN components whose parallel execution is coordinated by the object manager. Both Manager, collector and stages are included in the definition of a Parallel Object (PO), [6]. Applications that deploy the PO pattern can exploit the inter-object parallelism as much as the internal or intra-object parallelism. A PO instance object has a similar structure to that of an object in Smalltalk, and additionally defines as cheduling politics, previously determined that specifies the way in which one or more operations carried out by the instance synchronize [6], [8]. Synchronization policies are expressed in terms of restrictions; for instance, mutual exclusion in reader/writer processes or the maximum parallelism allowed for writer processes. Thus, all the parallel objects derive from the classic definition of a class more synchronization restrictions (mutual exclusion and maximum parallelism), which are now included in that definition [3]. Objects of the same class share the specification contained in the class of which are instances. The inheritance allows objects to derive a new specification from the one that already exists in the super-class. Parallel objects support multiple inheritance in the CPAN model. With the strategy divide and conquer as CPAN the parallel processing technique called n-Tree is used to parallelize sequential code that solve classic problems that can be partitioned by divide and conquer a n-ary tree such as sum of numbers, ordering of numbers and N-body problem. It shows the performance analysis of these implementations (speedup, cpi, etc.), comparing them with their corresponding sequential implementations to demonstrate their usefulness: programmability and performance.

Keywords. CPANS, parallel objects, communication patterns, divide and conquer, n-tree, structured parallel programming.

1 Introduction

At moment the construction of concurrent and parallel systems has less restraints than ever, since the existence of parallel computation systems, more and more affordable, of high performance, or HPC (High Performance Computing) has brought to reality the possibility of obtaining a great efficiency in data processing without a great rise in prices. Even though, open problems that motivate research in this area still exist. Some of this problems of parallel programming environments amount to the users acceptance, which usually depends on whether they can offer complete expressions of the parallel programs behaviour that are built with these environments [7]. At the moment in OO application systems, the scientific community interested in the study of concurrency only accepts standards for programming environments based on Parallel Objects (POs). A first approach that tries to tackle this problem is to let the programmer to develop his programs according to a sequential programming style, then, he can automatically obtain the parallelised parts of the code with the help of a specific environment.

However, intrinsic implementation difficulties exist mainly due to the difficult definition of programming languages formal semantics that refrain from the automatic (without user participation) sequential code parallelisation, and thus the problem of generating parallelism in an automatic way for a general application continues unsolved. The so called structured parallelism has become a promising approach to solve the mentioned problem. In general, parallel applications follow predetermined patterns of execution. Communication patterns are rarely arbitrary and are not structured in their logic [10]. We are interested, in particular, to do research work that has to do with parallel applications that use predetermined communication patterns, among other componentsoftware. Even so, with this promising approach, at least the following ones have currently been identified as important open problems: The lack of acceptance structured parallel programming environments of use to develop applications, [2], The necessity to have patterns or High Level Parallel Compositions, the Determination of a complete set of patterns as well as of their semantics, [7], the adoption of an object-oriented approach, [6], [9]. CPANs are parallel patterns defined and logically structured that, once identified in terms of their components and of their communication, can be adopted in the practice and be available as high level abstractions in user applications within an OO-programming environment. The process interconnection structures of most common parallel execution patterns, such as pipelines, farms and trees can be built using CPANs, within the work environment of POs that is the one used to detail the structure of a CPAN implementation.

A structured approach to parallel programming is based on the use of communication/interaction patterns (pipelines, farms, trees, etc.), which are predefined

structures of users application processes. In such a situation, the structured parallelism approach provides the interaction-pattern abstraction and describes applications through CPANs, which are able to implement the patterns mentioned already. The encapsulation of a CPAN should follow the modularity principle and it should provide a base to obtain an effective reusability of the parallel behaviour to be implemented. When there is the possibility of attaining this, a generic parallel pattern is built, which in its turn provides a possible implementation of the interaction structure between processes of the application, independently of the functionality of these. In addition, it is in line with the structured approach we have adopted that is the enrichment of traditional parallel environments with libraries of program skeletons [9] that concrete communication patterns represent. What it really means is a new design approach to parallel applications. Instead of programming a concurrent application from the beginning and controlling the creation of processes as well as the communications among them, the user simply identifies those CPANs that can implement the adapted patterns to the communication needs of his application and uses them together with the sequential code that implements the computations that individually carry out their processes. Several significant and reusable parallel patterns of interconnection can be identified in multiple applications and parallel algorithms which has resulted in a wide library of communication patterns between concurrent processes such as CPANs whose details are found in [14] and [15]. In the present work we have implemented the partitioning strategy divide and conquer using N-Tree pattern as a generic CPAN and using the object orientation paradigm we have realized its concretion in three particular applications: the add of numbers, the sorting of numbers and the solution of N-body particles, using for this the choice of three different strategies for the parallel implementation as CPANs of its sequential algorithms. In this way it is the user's own applications that specify the semantics of the N-Tree-Divide and Conquer according to the requirements of the software that was developed. Finally we show an analysis of the performance in terms of acceleration Amdhal refers to the Cpans TreeDV in solving the above problems, for a restricted range of exclusive processors in a parallel computer.

2 High Level Parallel Compositions (CPAN)

A CPAN comes from the composition of a set three object types: An object manager that represents the CPAN itself and makes an encapsulated abstraction out of it that hides the internal structure. The object manager controls a set of objects references, which address the object Collector and several Stage objects and represent the CPAN components whose parallel execution is coordinated by the object manager.

The objects Stage are objects of a specific purpose, in charge of encapsulating an client-server type interface that settles down between the manager and the slave-objects. These objects do not actively participate in the composition of the CPAN, but are considered external entities that contain the sequential algorithm that constitutes the solution of a given problem. Additionally, they provide the

necessary inter-connection to implement the semantics of the communication pattern which definition is sought. In other words, each stage should act a node of the graph representing the pattern that operates in parallel with the other nodes. Depending on the particular pattern that the implemented CPAN follows, any stage of it can be directly connected to the manager and/or to the other component stages.

The Collector object we can see an object in charge of storing the results received from the stage objects to which is connected, in parallel with other objects of CPAN composition. That is to say, during a service request the control flow within the stages of a CPAN depends on the implemented communication pattern. When the composition finishes its execution, the result does not return to the manager directly, but rather to an instance of the Collector class that is in charge of storing these results and sending them to the manager, which will finally send the results to the environment, which in its turn sends them to a collector object as soon as they arrive, without being necessary to wait for all the results that are being obtained. In summary, a CPAN is composed of an object manager that represents the CPAN itself, some stage objects and an object of the class Collector, for each petition that should be managed within the CPAN. Also, for each stage, a slave object will be in charge of implementing the necessary functionalities to solve the sequential version of the problem being solved (Figure 1). For details see [14].

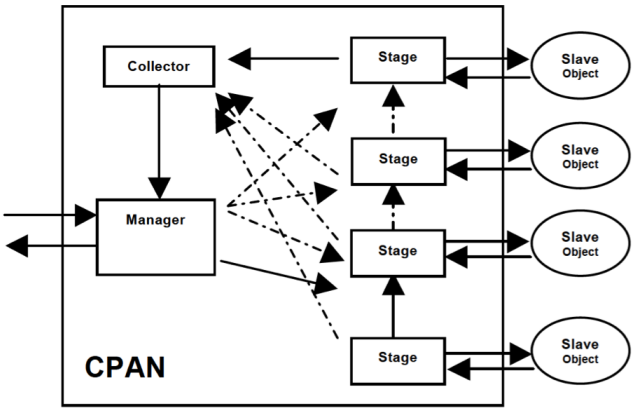

Fig. 1. Internal structure of CPAN. Composition of its components.

The Figure 1 shows the pattern CPAN in general, without defining any explicit parallel communication pattern. The box that includes the components, represents the encapsulated CPAN, internal boxes represent compound objects (collector, manager and objects stages), as long as the circles are the objects slaves associated to the stages. The continuous lines within the CPAN suppose that at least a connection should exist between the manager and some of the

component stages. Same thing happens between the stages and the collector. The dotted lines mean more than one connection among components of the CPAN.

2.1 The CPAN seen as Composition of Parallel Objects

Manager, collector and stages are included in the definition of a Parallel Object (PO), [6]. Parallel Objects are active objects, which is equivalent to say that these objects have intrinsic execution capability, [6]. Applications that deploy the PO pattern can exploit the inter-object parallelism as much as the internal or intra-object parallelism. A PO-instance object has a similar structure to that of an object in Smalltalk, and additionally defines a scheduling politics, previously determined that specifies the way in which one or more operations carried out by the instance synchronize, [6], [8]. Synchronization policies are expressed in terms of restrictions; for instance, mutual exclusion in reader/writer processes or the maximum parallelism allowed for writer processes. Thus, all the parallel objects derive from the classic definition of a class plus the synchronization restrictions (mutual exclusion and maximum parallelism), which are now included in that definition [3]. Objects of the same class share the specification contained in the class of which are instances. The inheritance allows objects to derive a new specification from the one that already exists in the super-class. Parallel objects support multiple inheritance in the CPAN model.

2.2 Communication Types in the Parallel Objects of CPAN

Parallel objects define 3 communication modes: synchronous, asynchronous communication and synchronous future communication.

1. The synchronous communication mode stops the client activity until it receives the answer of its request from the active server object [1].
2. The asynchronous communication does not delay the client activity. The client simply sends the request to the active object server and its execution continues afterwards [1].
3. The asynchronous future will delay client activity when the method's result is reached in the client's code to evaluate an expression. For details see [11].

The asynchronous and asynchronous future communication modes carry out the inter-objects parallelism by executing the client and server objects at the same time.

2.3 The Synchronization Restrictions of a CPAN

It is necessary to have synchronization mechanisms available when parallel request of service take place in a CPAN, so that the objects that conform it can negotiate several execution flows concurrently and, at the same time, guarantee the consistency in the data that being processed. Within any CPAN the restrictions MAXPAR, MUTEX and SYNC can be used for correct programming of their methods.

1. MAXPAR: The maximum parallelism or MaxPar is the maximum number of processes that can be executed at the same time. That is to say the MAXPAR applied to a function represents the maximum number of processes that can execute that function concurrently.
2. MUTEX: The restriction of synchronization mutex carries out a mutual exclusion among processes that want to access to a shared object. The mutex preserves critical sections of code and obtains exclusive access to the resources.
3. SYNC: The restriction SYNC is not more than a producer/consumer type of synchronization.

The details of the algorithms and their implementation can be seen in [14] and [15].

3 Construction of a CPAN

Each CPAN is made up of several objects: an object manager, some stage objects and a collector object for each request sent by client objects of the CPAN. In PO the necessary base classes to define the manager, collector, stages objects that compose a CPAN - the implementation details are in [14] - are the next ones: Abstract class ComponentManager,

Abstract class ComponentStage and Concrete class ComponentCollector. With the base-classes of the PO model of programming, it is now possible to build concrete CPANs. To build a CPAN, first it should have made clear the parallel behavior that the user application needs to implement, so that the CPAN becomes this pattern itself. Several parallel patterns of interaction have long been identified in Parallel Programming, such as farms, pipes, trees, cubes, meshes, a matrix of processes, etc. Once identified the parallel behavior, the second step consists of elaborating a graph of its representation, as an informal design of the objective system. This practice is also good for illustrating the general characteristics of the desired system and will allow us to define its representation with CPANs later on, by following the pattern proposed in the previous section. When the model of a CPAN has already been made clear, it defines a specific parallel pattern; let's say, for example, a tree, or some other mentioned pattern, and then the following step will be to do its syntactic definition and specify its semantics.

Finally, the syntactic definition prior to any programmed CPAN is transformed into the most appropriate programming environment, with the objective of producing its parallel implementation. It must be verified that the resulting semantics is the correct one. To attain this, we use several different examples to demonstrate the generality and flexibility of the application CPAN-based design and the expected performance and quality as a software component. Some support from an integrated development environment (IDE) for Parallel Programming should be provided in order to validate the component satisfactorily. The parallel patterns worked in the present investigation have been the pipeline and the binary-tree to solve the sorting problem using two different algorithms.

4 The Technique of Divide and Conquer as a CPAN

The technique of it Divide and Conquer it is characterized by the division of a problem in sub-problems that have the same form that the complete problem [4]. The division of the problem in smaller sub-problems is carried out using the recursion. The method recursive continues dividing the problem until the parts divided can no longer follow dividing itself, and then they combine the partial results of each sub-problem to obtain at the end the solution to the initial problem [4]. In this technique the division of the problem is always made in two parts, therefore a formulation recursive of the method Divide and Conquer form a binary tree whose nodes will be processors, processes or threads [5], [12].

4.1 Representation of the Tree - Divide and Conquer (TreeDV) as a CPAN

The representation of the patron tree that defines the technique of it Divide and Conquer as CPAN has their model represented in figure 2. This parallel solution offers the prospect of traversing several parts of the tree simultaneously in the Cpan TreeDV. Once a division is made into two parts, both parts can be processed simultaneously executing the sequential algorithm contained in the slave object associated to the nodes of the tree. Though a recursive parallel solution could be formulated. One could simply assign one process o thread to each node in the tree.

4.2 Use and Utility of CPAN TreeDV

The potential of the CPAN TreeDV in the solution of various problems that can be solved by applying the technique of divide and conquer generating a binary tree is shown below.

Adding a list of numbers: A recursive definition for adding a list of numbers is:

```
int add(int *s) {
   if (number(s) <= 2) then return (n1+n2);
   else{
         divide(s, s1, s2);
         part_sum1= add(s1);
         part_sum2= add(s2);
         return (part_sum1+part_sum2);
        } }
```

In the code, $number(s)$ returns the number of numbers in the list pointed to by s. If there are two numbers in the list, they are called $n1$ and $n2$. If there is one number in the list, it is called $n1$ and $n2$ is zero. If there are no numbers,

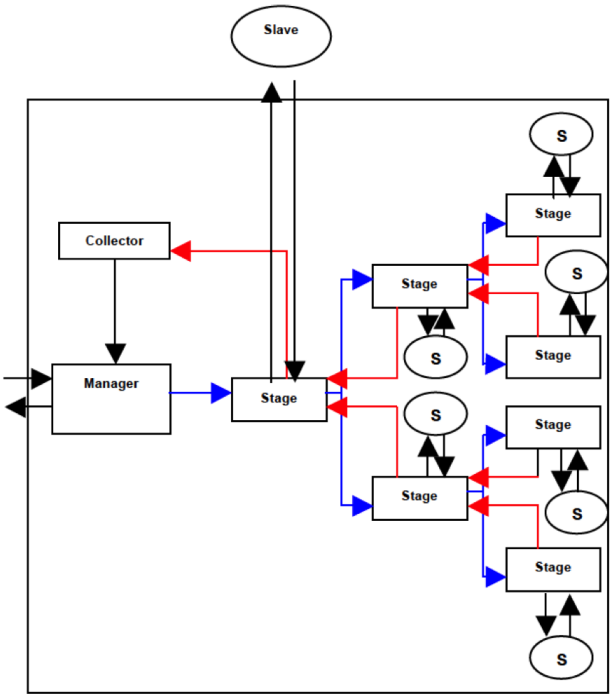

Fig. 2. The Cpan of a TreeDV.

both $n1$ and $n2$ are zero. Separate if statements could be used for each of the cases; 0,1, or 2 numbers in the list. Each would cause termination of the recursive call, [16].

Our parallel proposal is to make use of the Cpan TreeDV. The nodes of the binary tree in the CPAN (stage processes) will be created dynamically through the execution of the proposed sequential algorithm and that is associated to the slave objects of each node in the tree. A more efficient solution adopted is to reuse stage process at each level of the tree, ie the combining act of summation of the partial sums can be done as ilustrated in figure 3. Once the partial sums have been formed, each odd-numbered stage process passes its partial sum to the adjacent even-numbered stage process, that is, $Stage1$ passes its sum to $Stage0$, $Stage3$ to $Stage2$, $Stage5$ to $Stage4$, and so on. The even-numbered stage processes then add the partial sum with its own partial sum and pass the result onward, as shown in figure 3. This continues until $Stage0$ has the final result which is passed to the Collector object of the CPAN and this in turn sends it to the Manager object that passes the result to the user.

Quicksort sorting algorithm: The Quicksort sorting was created by Hoare and is based on the paradigm of divide and conquer. As a first step the algorithm

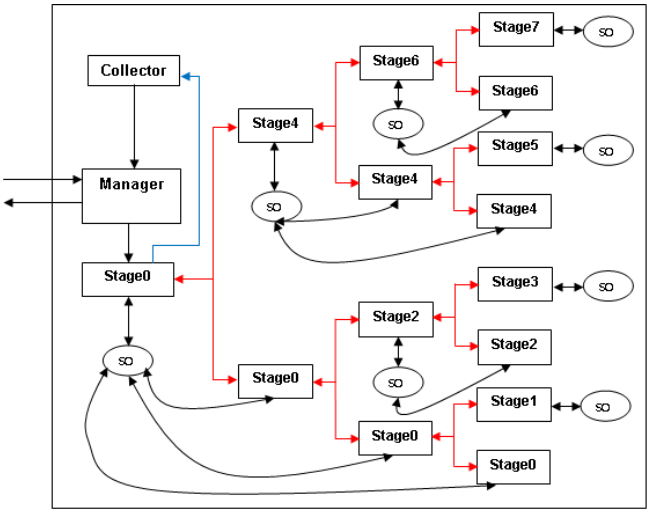

Fig. 3. Adding a list of numbers using CPAN TreeDV.

selects as a pivot one of the elements of the data set you have to order. The array is then partitioned on either side of the pivot: elements are moved so that those greater than the pivot are to its right, whereas the others are to its left. If now the sections of the array on either side of the pivot are sorted independently by recursive and parallel calls of the algorithm [4], in this case through the stage TreeDV CPAN objects, the final result is a completely sorted array, no subsequent merge step being necessary.

```
Algorithm QuickSort(T[,..j]) {
    var l;
     if (j-i  is sufficiently small) then insert(T[i..j])
     else {
           l= pivot(T[i..j]);
           QuickSort(T[i..l-1]);
           QuickSort(T[l+1..j]);
          } }
```

To balance the sizes of the two subinstances to be sorted, we would like to use the median element as the pivot. Unfortunately, finding the median takes more time it is worth. For this reason we simply use an arbitrary element of the array as the pivot, hoping for the best.

```
Algorithm pivot(T[i..j]) {
    var l;
    p=T[i];  k=i;  l=j+1;
    repeat {k=k+1;} until ((T[k]>p) or (k>=j));
    repeat {l=l-1;} until (T[l]<=p);
```

```
while (k<l)
   {
     swap(T[k],T[l]);
     repeat {k=k+1;} until (T[k]>p);
     repeat {l=l-1;} until (T[l]<=p);
   }
swap(T[i],T[l]);
return l; }
```

Suppose subarray T[i..j] is to be pivoted around p=T[i]. One good way of pivoting consists of scanning the subarray just once, but starting at both ends. Pointers k and l are initialized to i and j+1, respectively. Pointer k is then incremented until T[k]¿p, and pointer I is decremented until T[l]¡=p. Now T[k] and T[l] are interchanged. This process continues as long as k¡l. Finally, T[i] and T[l] are interchanged to put the pivot in its correct position [4], (see figure 4) .

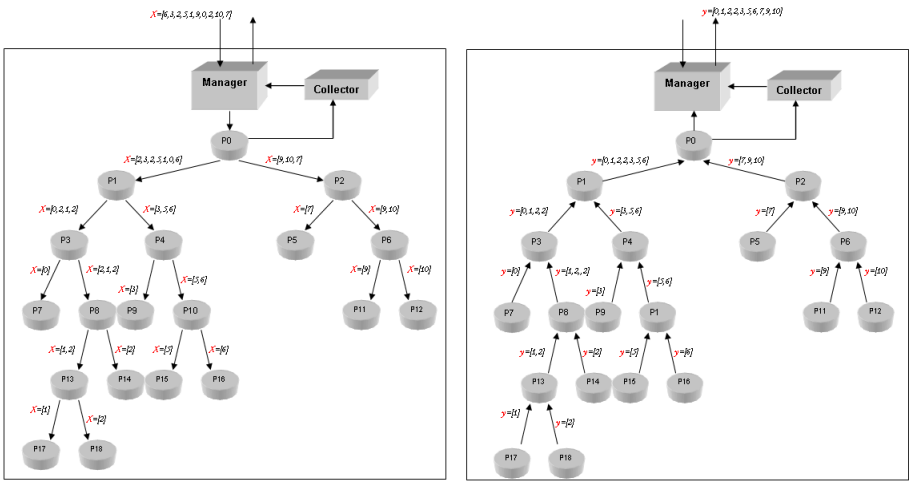

Fig. 4. Sequence of the QuickSort sort algorithm using CPAN TreeDV.

N-Body Problem: The N-Body problem is concerned with determining the effects of forces between bodies, por example, astronomical bodies that are attracted to each other through gravitational forces or charged particles are also influenced by each other according to electrostatic law. We provide the basic equations to enable the application to be coded as a CPAN TreeDV using as a case study the N-Body problem in terms of particles charged according to Coloumb's electrostatic law; particles of opposite charge are attracted and those of like charge are repelled. Also charged particles may move away from each other. The objective is to find the positions and movements of the particles

in the space that are subject to electrostatic forces from other particles using Coulomb laws.

For a computer simulation, we use values at particular times, t_0, t_1, t_2, etc., the time intervals being as short as posible to achieve the most accurate solution. Let the time interval be Δt. Then, for a particular particle of mass m, the force is given by:

$$F = \frac{m(v^{t+1} - v^t)}{\Delta t},$$

and a new velocity by:

$$v^{t+1} = v^t + \frac{F\Delta t}{m},$$

where v^{t+1} is the velocity of the particle at time $t + 1$ and v^t is the velocity of the particle at time t. If a particle is moving at a velocity v over the time interval Δt, its position changes by:

$$x^{t+1} - x^t = v\Delta t,$$

where x^t is position at time t. Once particles move to new positions, the forces change and the computation has to be repeated. The computation of the attraction or not of N-particles according to their electrostatic charge is described in the following algorithm:

```
for (t=0; t< tmax; t++) {
   for(i=0;i<N;i++)
     {
      F=force(i);
      v[i]_new = v[i]+F*dt/m;
      x[i]_new = x[i]+v[i]_new*dt;
     }}

for(i=0;i<nmax;i++) {
   x[i]=x[i]_new;
   v[i]=v[i]_new;
  }
```

For each time period t, for each particle i, compute force on ith particle, compute new velocity and new position. For each particle i update velocity and position.

Parallelizing this algorithm can use partitioning where by groups of particles are the responsability of each process, and each force is carrried in distint messages between process. A large number of messages could result and it is not feasible if N is very large, [16]. The complexity can be reduced using the technique that a cluster of distant particles can be approximated as a single distant particle of the total mass of the cluster sited at the center of mass of the cluster.

This idea can be implemented as a CPAN by being applied recursively generating a m-ary tree, in particular way, a $quad-tree$ (a tree in which each node of tree has four children) based on the Barnes-Hut algorithm, as you can see in [13] and [16].

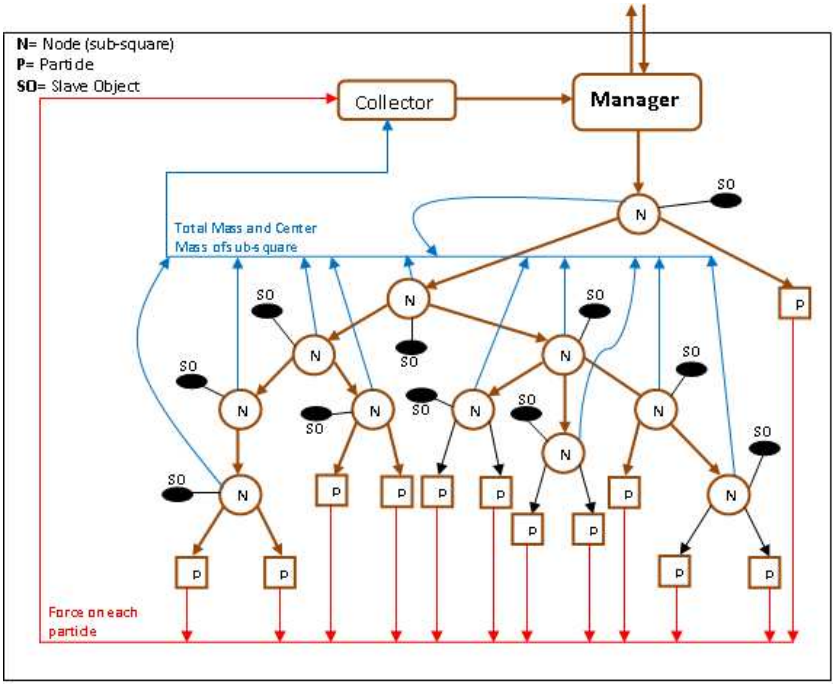

Fig. 5. Cpan QuadTree Particle.

A divide and Conquer formation to the problem using this clustering idea start for a two-dimensional space in which one square contains the particles. This square is recursively divided into four sub-squares creating a *quadtree*, ie a tree with up to four edges from each edge. If a sub-square contains no particles, the sub-square is deleted from futher considerarion. If a sub-square contains more than one particle, it is recursively divided until every sub-square contains one particle creates the *quadtree*. The tree will be unbalanced. The leaves represent cells each containing one particle.

The Figure 5 represents the resultant quadtree like a CPAN. In the "Cpan QuadTree Particle" of figure 5, the total mass and center of mass of the subsquare is stored at each node of tree. The force on each particle can be obtained by traversing the tree starting at the root, stopping at a node when the clustering approximation can be used for the particular particle, and otherwise continuing to traverse the tree downward.

5 Performance

Performance analysis of CPANS TreeDV for adding a list of numbers, sorting a list of numbers using Quicksort algorithm and the N-Body Problem are shown. The aim is to show that, at least for these problems, the performances obtained are "good" based on the model of the CPAN. The CPAN TreeDV performance to solve the problems mentioned was carried out on a parallel computer with 64 processors, 8 GB of main memory, high-speed buses and distributed shared memory architecture. Performance measures obtained in implementing the CPANs TreeDV using Divide and Conquer Technique is carried out with the following restrictions execution:

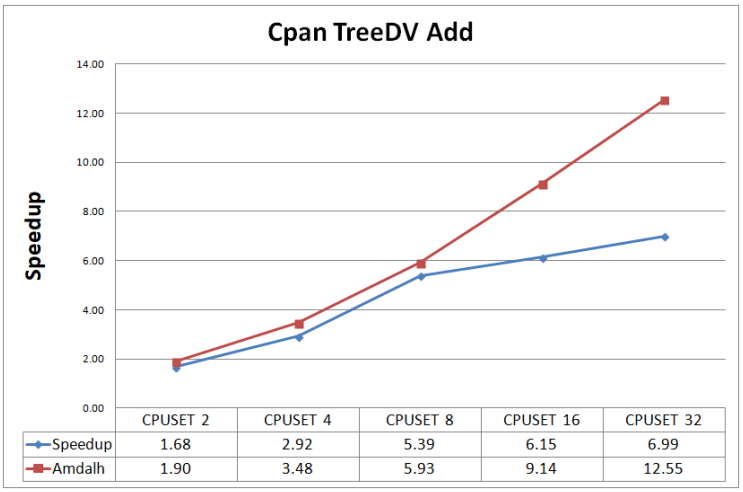

Fig. 6. Speedup scalability found for CpanTreeDV in solution of adding numbers problem for 2, 4, 8, 16 and 32 exclusive processors.

- In the adding numbers problem, the same sequential sum algorithm was used in each of the slave objects associated with the stages (nodes) of binary tree that is generated in the Cpan TreeDV (see figure 3).
- In the sorting numbers problem, parallel implementation of sequential sorting algorithm based on a CPAN TreeDV is Quicksort sorting algorithm based on a binary tree (see figure 4).
- In both cases, both in the adding numbers problem and in the sorting numbers problem, 50000 random integers were generated; cach number generated in a range between 0 and 50000, allowing make a sufficient charge for processors and thereby observe the performance improvement CPAN TreeDV.
- In the N-body problem, we work with a simulation of 50000 particles with electrostatic charge moving randomly in space. The calculations were: find

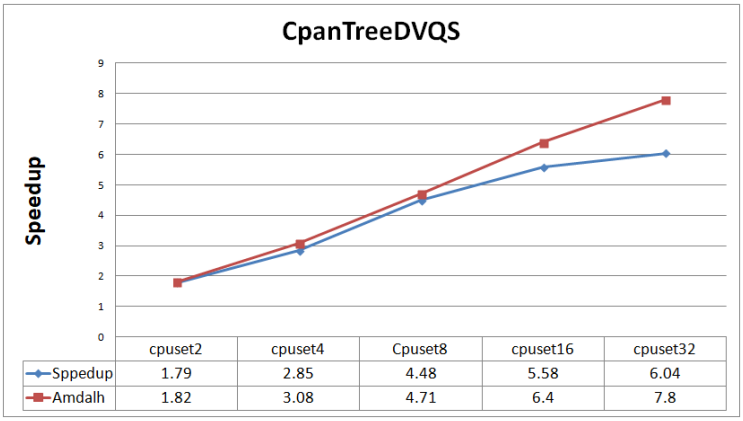

Fig. 7. Speedup scalability found for CpanTreeDVQS in solution of sorting numbers problem for 2, 4, 8, 16 and 32 exclusive processors.

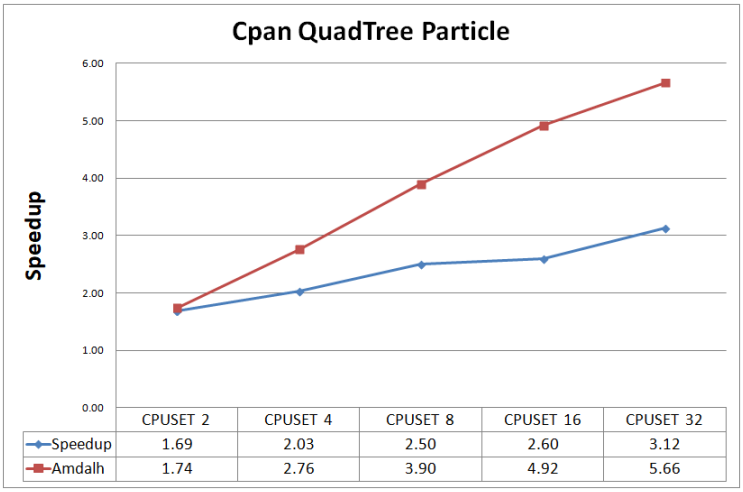

Fig. 8. Speedup scalability found for CpanQuadTree in solution of N-body problem for 2, 4, 8, 16 and 32 exclusive processors.

the positions and movements of the particles in the space that are subject to electrostatic forces from other particles using Coulomb laws. For this, the CPAN QuadTree calculated with the sequential algorithms associated with the slave objects of the generated M-tree, the masses and the forces of each particle (see figure 5).

These execution conditions allow a sufficient load for the processors and show the good performance of the Cpan TreeDV when solving them. For all of them the execution was performed in 2, 4, 8, 16 and 32 exclusive processors and the

results are shown in the figure 6, figure 7 and figure 8. In them show the series of measurements obtained including their corresponding sequential versions for Cpans TreeDV, TreeDVQS and QuadTree, magnitude speedup found and the upper bound on the magnitude of speedup using for that Amdahl's law, moreover the runtime execution in seconds of the programs.

Parallel executions of CPANS have a time shorter than the time used by their corresponding sequential versions, as expected. The execution times of their parallel versions CPANS improve as the number of processors is increased, ie, as is increasing the number of processors with which CPANS are executed, their execution times are decreasing. A value of the magnitude called speedup is appreciated ever upward on improving execution times of parallel CPANS respect to its sequential counterpart, but always below the levels of Amdahl's Law calculated, obtaining "good" yields.

6 Conclusions

We have presented a method for design of concurrent applications based on the construction of High Level Parallel Compositions or CPANS and which are usually used in different platforms, such as C ++ and POSIX Threads. We discuss the implementation of CPANs treeDV as generic and reusable patterns of communication/interaction between processes which implements the algorithm design technique called divide and conquer making use of an N-tree as a pattern of communication associated, which can even be used by inexperienced parallel application programmers to obtain efficient code by only programming the sequential parts of their applications.The CPAN TreeDV has been reused in the communication/interaction between the processes of three solved problems with different implementation strategies of their respective sequential algorithms of solution: the adding numbers problem, the sorting numbers problem and de N-body (particles) problem. This selected problems have been included to show speedup and low execution times about their best sequential version of the algorithms that solve these problems. We have also obtained good performance in their executions and speedup scalability compared to Amdahls law on the number of processors used to obtain the solution.

References

1. Andrews, G.R.: Foundations of Multithreaded, Parallel, and Distributed Programming. Addison Wesley (2000)
2. Bacci, B., Danelutto, M., Pelagattii, S., Vaneschi, M.: SklE: A Heterogeneous Environment for HPC Applications Parallel Computing. Springer, Vol. 25, No. 13–14 (1999)
3. Birrell, B.: An Introduction to Programming with Threads. Digital Equipment Corporation, Systems Research Center, Palo Alto California, USA (1989)
4. Brassard, G., Bratley, P.: Fundamentals of Algorithmics. Prentice-Hall (1997)

5. Brinch, H.: Model Programs for Computational Science. A programming methodology for multicomputers, Concurrency, Practice and Experience, Volume 5, Number 5 (1993)
6. Corradi, A., Leonardi, L.: PO constraints as tools to synchronize active objects. Journal Object Oriented Programming, Vol. 4, No. 6, pp.41–53 (1991)
7. Corradi, A., Leonardo, L., Zambonelli, F.: Experiences toward an Object-Oriented Approach to Structured Parallel Programming. DEIS Technical Report No. DEISLIA-95-007 (1995)
8. Danelutto, M., Torquati, M: Loop parallelism: A new skeleton perspective on data parallel patterns. In Proceedings of Intl. Euromicro PDP: Parallel Distributed and Network-based Processing, Torino, Italy (2014)
9. Darlington, J.: Parallel programming using skeleton functions. In Proceedings PARLE93, Munich (1993)
10. Hansen, B.: Model programs for computational science: A programming methodology for multicomputers. Concurrency Practice and Experience, Vol. 5, No. 5 (1993)
11. Lavander, G., Kafura, D.: A Polymorphic Future and First-class Function Type for Concurrent Object-Oriented Programming in C++. Journal of Object-Oriented Systems (1995)
12. Liwu, L.: Java Data Structures and Programming. Springer Verlag, Germany (2002)
13. Roosta, S.: Parallel Processing and Parallel Algorithms. Theory and Computation, Springer (1999)
14. Rossainz, M., Capel, M.: A Parallel Programming Methodology using Communication Patterns named CPANS or Composition of Parallel Object. In Proceedings of 20TH European Modeling & Simulation Symposium, Campora S. Giovanni, Italy (2008)
15. Rossainz, M., Capel, M.: Design and implementation of communication patterns using parallel objects. International Journal of Simulation and Process Modelling, Volume 12, No. 1, Pp: 69–91 (2017)
16. Wilkinson, B., Allen, M.: Parallel Programming Techniques and Applications Using Networked Workstations and Parallel Computers. Prentice-Hall, USA (1999)

ISSN 1870-4069

Sistema educativo multimedia para el apoyo del aprendizaje autónomo de metodología de la programación

Eugenia Erica[1,2], Claudia Marina-Vicario[1,3], Yadira Navarro[2]

[1] Centro de Estudios Superiores en Educación (CESE), Ciudad de México

[2] Benemérita Universidad Autónoma de Puebla, México

[3] Instituto Politécnico Nacional (IPN), México

{eevclibra, ynavarro44}@gmail.com,
marina.vicario@gmail.com

Resumen. Es clara la diversidad de los alumnos en el área de programación, ya que así como hay estudiantes que de inmediato comprenden, hay estudiantes que pasa mucho tiempo para que puedan mostrar un ejercicio correcto. La solución es ser capaz de atender la diversidad, entonces el potencial de las TIC puede favorecer en diversas formas los procesos de autoaprendizaje y posibilitar distintas modalidades educativas. El propósito de este artículo es presentar el diseño y desarrollo de un Sistema educativo multimedia para el apoyo del aprendizaje autónomo de metodología de la programación (SEMMP), el cual considera las características del estudiante (perfil), para adaptar y presentar los contenidos basados en competencias para apoyar a desarrollar las habilidades en la programación en el tema de metodología de la programación. Finalmente se presentan los resultados obtenidos del sistema al realizar una prueba piloto con una muestra de estudiantes.

Palabras clave: Metodología de la programación, sistema educativo, multimedia, ingeniería web.

Multimedia Educational System for the Support of Autonomous Learning of Programming Methodology

Abstract. The diversity of the students in the programming area is clear, since just as there are students who immediately understand, there are students who spend a lot of time so that they can show a correct exercise. The solution is to be able to address diversity, then the potential of ICT can favor in different ways the processes of self-learning and enable different educational modalities. The purpose of this article is to present the design and development of a multimedia ed-

ucational system for the support of autonomous learning methodology programming (SEMMP), which considers the characteristics of the student (profile), to adapt and present content based on competences to support the development of programming skills in the subject of programming methodology. Finally, the results obtained from the system are presented when conducting a pilot test with a sample of students.

Keywords. Methodology of programming, educational system, multimedia, web engineering.

1. Introducción

En la carrera de Licenciatura en Ciencias en Computación de la Facultad de Ciencias de la Computación (FCC) de la BUAP se ha observado que los alumnos presentan algunas dificultades en la materia de Programación y en particular, con la metodología de la programación. Un fenómeno importante que se observa es que independientemente del profesor que imparte estos cursos introductorios de programación, existen estudiantes que han logrado aprender a programar y de la misma forma existen estudiantes que no logran entender los temas, que comúnmente son los que reprueban. Así, de los alumnos aprobados, en muchas ocasiones, sus calificaciones son muy bajas, lo que se refleja en la no acreditación de las siguientes materias que dan por sentados los conocimientos de las anteriores. Es evidente la existencia de una diferencia marcada entre cada profesor que imparte la materia de Metodología de la Programación, por lo que hemos afirmado que el estilo docente no se relaciona de manera significativa con el rendimiento académico de los alumnos, es por ello que resulta necesario apoyar al estudiante en el contexto del aprendizaje autónomo como proceso individual que depende del trabajo, de los hábitos de estudio y de las estrategias del alumno.

Por esta razón, el propósito de este trabajo es proporcionar un proceso de enseñanza que amplíe y mejore el repertorio de estrategias, para que cada alumno adelante el proceso de aprendizaje de manera autónoma, durante el semestre logrando un avance significativo en la materia de Metodología de la Programación y de esta manera disminuir los índices de reprobación.

El documento está conformado y organizado de la siguiente manera: en la sección 2, encontramos la fundamentación teórica del modelo de Bernad (ESEAC) con el que se basaron las estrategias de aprendizaje autónomo del Sistema educativo, la explicación del concepto de constructivismo, la conceptualización de aprendizaje autónomo, multimedia y perfil de ingreso académico; en la sección 3, se describe el diseño y desarrollo del Sistema educativo multimedia para el aprendizaje autónomo de la materia de MP; en la sección 3, encontramos las pruebas del sistema y finalmente en la sección 4, tenemos las conclusiones y trabajos futuros.

2. Marco teórico

En este apartado desarrollamos la fundamentación teórica que se utilizó para realizar el diseño y desarrollo del Sistema educativo multimedia para el apoyo del aprendizaje autónomo de metodología de la programación.

2.1. Dimensiones de la Escala de Estrategias de aprendizaje contextualizado (ESEAC)

Las estrategias de aprendizaje son entendidas como los procesos intencionales (conscientes) que permiten utilizar las estrategias cognitivas para alcanzar una determinada meta o tarea de aprendizaje, de esta forma el estudiante lleva a cabo un conjunto de operaciones mentales en una secuencia determinada [1]. Algunas estrategias de aprendizaje son: abstracción, mapa conceptual y resolución de problemas, por mencionar algunos ejemplos.

Las estrategias de aprendizaje contextualizado de Bernad trabajan con la parte: cognitiva, estratégica y contextualizada, aunque su objetivo es proporcionar al profesor criterios e instrumentos que le sirvan para conocer adecuadamente lo que sus alumnos han aprendido. En nuestro estudio recuperamos las dimensiones del aprendizaje contextualizado para desarrollar nuestras estrategias de aprendizaje autónomo que se pueden consultar en [2], en donde la actividad es una actividad de utilidad, para esto consideramos:

1. La intencionalidad o dimensión "estratégica" del proceso de aprender.
2. La contextualización del aprendizaje. Un modelo que se ocupa del aprendizaje referido a un marco concreto es el que se impone en la actualidad.

Primer factor de contextualización: el nivel académico del alumno.

El constructivismo clásico piagetano, debidamente traducido en sus correlatos cognitivos y completado con las aportaciones de los modelos de procesamiento de información, proporciona una directriz básicamente esclarecedora que ayuda a determinar "que se puede exigir" y, por tanto, "de que hay que evaluar" al alumno según se encuentre en uno y otro nivel académico [3]. El nivel de dominio, en estudios superiores es concreto y formal al 50%.

Las estructuras de Piaget - pensamiento sensomotor, preoperacional, concreto y formal- equivalen a macro estrategias que por sí solas sirven a modo de hitos psicológicamente significativos para esclarecer y dirigir el proceso de aprender.

Segundo factor de contextualización: especificidad de los contenidos curriculares.

El modelo consta de dos secciones, la que se ocupa de definir cómo procesa el escolar la información relativa a las materias que estudia (estrategias de procesamiento), y la que intenta calibrar el control que el aprendiz ejerce sobre su propio modo de regularse en la realización de los procesos incluidos en la sección primera (estrategias de autocontrol o de apoyo) [3].

El modelo se fundamenta en la concepción del aprendizaje cognitivo-constructivista y se desglosa en 7 dimensiones del proceso de aprender, 8 estrategias y 3 niveles de ejecución [3].

2.2. Paradigma constructivista

En trabajos de García y Gómez se afirma que el remoto autor del constructivismo sería Lao-tsé, quien ya en el siglo VI A. C. decía..."Id donde esté la gente. Aprended de ella. Mostradle su amor. Partid de lo que ya sabe. Construid sobre lo que ya han hecho. Y cuando hayáis terminado vuestra tarea, sabremos que hemos sido exitosos cuando ellos digan: lo hicimos nosotros mismos" [4]. Posteriormente Sócrates inicia su metodología "casi constructivista". Aristóteles concreta esta posición mostrando que los pensamientos se crean a partir de los hechos observados en la existencia, se accede a una combinación de rasgos o atributos que podemos absorber teóricamente.

El enfoque constructivista entró en Latinoamérica con publicaciones sobre el constructivismo en diferentes volúmenes, diferentes países plantearon en sus reformas que su currículo debía ser constructivista, capacitando a su personal. La fuerza de este enfoque surge de las ideas de Piaget, Brunner, Gagné, Vigotsky, Ausebel, de los trabajos de Coll en 1985-1989 y de la reforma curricular establecida en España en 1989 [5]. Este enfoque consiste en que la enseñanza se forma por el alumno en la participación continúa con los objetos en el conocimiento, intercambios sociales y la motivación. El conocimiento parte de la idea previa en el origen para el conocimiento nuevo, siendo este esencialmente activo. Entonces podemos entender el constructivismo como un conjunto de concepciones sobre el aprendizaje, que provienen de dos teorías básicas del desarrollo cognoscitivo [6,7,8].

El aprendizaje en el enfoque constructivista no se resume en la transmisión y acumulación de conocimiento sino trata de una construcción que hace el alumno de él mismo en la realidad, se trata de un proceso activo para edificar en el conocimiento en la experiencia y la información que se recibe.

2.3. Aprendizaje autónomo

El aprendizaje autónomo es un aprendizaje estratégico en el que la persona toma decisiones claves sobre su propio aprendizaje: autodirigiéndolo en función de unas necesidades, metas o propósitos, auto regulándolo (seleccionando alternativas, acciones, tiempos) y autoevaluándolo, de acuerdo con los recursos y escenarios de que dispone y de las exigencias y condiciones del contexto. Con el aprendizaje autónomo la persona aprende a aprender gracias al entrenamiento y desarrollo de competencias o habilidades cognitivas, afectivas e interactivas, pero también, y de manera esencial, gracias al desarrollo de habilidades metacognitivas [9].

Para que los estudiantes logren aprender a aprender, es necesario que se les enseñe a incorporar estrategias de aprendizaje, concientizarlos sobre la forma de cómo aprenden.

El desarrollo de procesos de aprendizaje autónomo exige: Planeación, organización, acción, constancia en la acción, disciplina, sistematización, hábito de trabajo, búsqueda de alternativas de solución de problemas, creatividad y desarrollar una enseñanza estratégica [10]. Una de las relaciones más cercanas del aprendizaje autónomo lo acerca a otros tipos de aprendizaje como el aprendizaje activo, el aprendizaje basado en problemas, el colaborativo y el cooperativo.

2.4. Multimedia

En el ámbito de la computación el término multimedia designa el uso de varios recursos o medios, como audio, video, animaciones, texto y gráficas en una computadora. Sin quedarse, sólo, en un collage de medios, al integrar los datos que puede manejar la computadora, la multimedia ofrece posibilidades de creatividad mediante los sistemas de computación [11].

La tecnología de multimedia toma auge en los video-juegos, a partir de 1992, cuando se integran: audio (música, sonido estéreo y voz), video, gráficas, animación y texto al mismo tiempo. La principal idea multimedia desarrollada en los video juegos es: que se pueda navegar y buscar la información que se desea sobre un tema, sin tener que recorrer todo el programa, que se pueda interactuar con la computadora y que la información no sea lineal sino asociativa [11].

2.5. Perfil de ingreso académico

Un perfil de ingreso es la caracterización posible de los estudiantes a partir de sus comportamientos académicos previos y del establecimiento de sus conductas sociodemográficas [12]. La elaboración de un perfil permite registrar las características de una determinada población o de una parte de tal población de forma que se posibilite una aproximación curricular, evaluativa y de seguimiento de aquellos factores que pueden estar influyendo en el desarrollo de la trayectoria escolar de la misma.

El perfil identificado debe permitir una descripción global e integradora de un número amplio de variables relacionadas no sólo con características socioeconómicas y demográficas sino igualmente con el desarrollo y funcionamiento de habilidades cognitivas y su relación con otros procesos [13].

La integralidad del diagnóstico condujo a la identificación de una serie de variables a estimar dadas las características de la Facultad de Ciencias de la Computación y de acuerdo con la investigación realizada sobre la reprobación, se seleccionaron seis variables principales con sus respectivas dimensiones, a saber: razonamiento científico, decisión vocacional, habilidad lecto-comprensiva, hábitos de estudio, estilos de aprendizaje y administración de tiempo. Cada una de estas variables fue estimada con base en la aplicación en línea de instrumentos adaptados para su uso a través de este medio. Se emplearon instrumentos validados en sus versiones originales y, en su mayoría, de conocimiento y uso común por investigadores y docentes.

En nuestro estudio se considera este perfil académico para nutrir como insumo de información al sistema educativo para el aprendizaje autónomo de la materia de Metodología de la Programación. Este sistema necesita estar bien informado de las características académicas de los estudiantes que se registren, para la selección de actividades de aprendizaje.

3. Desarrollo y pruebas del sistema educativo multimedia para el apoyo del aprendizaje autónomo de metodología de la programación

El sistema se desarrolló bajo el modelo cliente-servidor donde todo comienza con una petición o requerimiento HTTP iniciado en un navegador por un cliente que quiere acceder a un recurso de nuestro sitio web por medio de una dirección URL. La dirección URL apunta a la localización física de una página de extensión .ASP.

3.1. Actividades de aprendizaje autónomo

Las actividades de aprendizaje autónomo son acciones específicas que facilitan la ejecución de la estrategia de aprendizaje autónomo. Estas actividades hacen énfasis en el proceso intelectual que se encuentra especificado para cada idea principal y además son la base del diagnóstico académico así que las preguntas se presentan de la manera más clara posible, para que el alumno no caiga en confusiones. En cada actividad se le presenta al alumno la información del tema y subtema que está cursando.

Se diseñaron cinco tipos de actividades: Booleana, Opción Múltiple, Completar Texto, Relacionar y Arrastrar. En cada actividad es posible que el alumno se encuentre con un poco de indecisión al momento de responder a su actividad, en este caso todas las actividades cuentan con la opción de regresar a la clase que observaron antes de dicha actividad, como ayuda para resolver dudas, sin otorgar directamente la respuesta.

Es importante no carecer de información, así como tampoco disponer de información excesiva que confunde y consume tiempo. Por esto es necesario resaltar los aprendizajes relevantes de cada unidad, es decir diferenciar un aprendizaje importante de uno secundario.

Para definir cuáles son los aprendizajes relevantes de las unidades, temas o áreas que se seleccionaron para ser diagnosticadas, se tiene el siguiente método que consta de tres pasos [14].

a) Seleccionar las ideas principales o esenciales de cada una de las unidades, temas o áreas elegidas.
b) Determinar los procesos cognoscitivos deseables para cada idea esencial. Una vez que se seleccionaron las ideas principales del contenido de la enseñanza, se realiza un análisis de cada una de ellas para determinar los procesos intelectuales que el alumno pondrá en juego al haberlas aprendido.

   Los procesos intelectuales característicos del aprendizaje escolar utilizados son [14]: Conocimiento (repetir y reconocer información), comprensión (ejemplificar, comparar e interpretar la idea esencial), aplicación (integrar varias ideas, aplicar la idea esencial en situaciones nuevas o en la solución del problema), análisis (analizar la idea principal), síntesis y evaluación cognitivo.

   Al final obtendremos una puntuación de las ideas principales (conceptos, hechos, procedimientos), junto con el señalamiento de los procesos cognitivos que se esperan para cada uno de ellos.

c) Señalar el tipo de aprendizaje referido en cada idea esencial, en función al programa de estudios. El aprendizaje se clasifica en diferentes formas. Quesada propone hacerlo en tres categorías: indispensable, esencial y antecedente [14].
Los aprendizajes relevantes conforman parte del perfil académico deseado del alumno. Una vez que se tiene definido por completo el perfil académico deseado (aprendizajes relevantes) del alumno, así como sus antecedentes, se comienza la elaboración del instrumento diagnóstico, mediante el cual se le evaluará.

3.2. Desarrollo del Sistema educativo

En la Fig. 1 se observa que el usuario tiene acceso al Sistema educativo a través de links que le permitirán irse desplazando por todo el Sistema.

Fig. 1. Entrada al Sistema educativo para el aprendizaje de la materia de MP.

Fig. 2. Dentro del aula en el Sistema educativo.

Una vez dentro de la escuela, se muestran vínculos, a los que se podrán acceder a: Aula, Cafetería, Biblioteca, secretaria, Tablón de anuncios, Correo, Tareas Programadas, Ayuda y desplazarse por el sitio. En la sección Aula encontraremos foros, contenidos de la materia, ideas clave, herramientas multimedia, actividades de aprendizaje, un archivador, y tablón de anuncios (ver Fig. 2).

Una vez que seleccionamos la materia y el tema, el sistema pide que se identifique el estudiante, mediante una clave de acceso.

El Sistema educativo tiene programadas las clases de todo el curso (ver Fig. 3), permitiendo al alumno seleccionar la clase de aprendizaje que desea aprender o repasar.

Fig. 3. Pantalla de las clases de aprendizaje de la materia de MP en el Sistema educativo.

El Sistema educativo tiene programadas las actividades en ambientes virtuales para el aprendizaje (ver Fig. 4), en donde los estudiantes reciban una instrucción y al mismo tiempo son partícipes de ella jugando un rol activo para fomentar la colaboración de los mismos y enriqueciendo el acervo que el sistema ofrezca para dicho aprendizaje.

Fig. 4. Pantalla de una clase de aprendizaje y su actividad de aprendizaje de la materia de MP en el Sistema educativo.

En la Fig. 5 se observa una actividad de relacionar el concepto con la imagen y en la Fig. 6 se observa una clase y su actividad de aprendizaje de opción múltiple.

En caso de que el alumno falle en su respuesta, se determinó que no se le puede decir literalmente que no tiene validez, así que en los dos tipos de respuesta, correcta o no, se guiará con un tip de apoyo para aclararle cual camino era el indicado, esto sin darles la respuesta, simplemente como una idea que despejará sus dudas y motivando a continuar con el curso.

Fig. 5. Pantalla de una clase de aprendizaje y su actividad de opción múltiple.

Se intenta provocar, para que cuestionen los conocimientos que ya tiene y se esfuerce por adquirir los que aún no domina. En el momento de contestar acertadamente el concepto, se le aumentará los puntos que le corresponden y se le motivará con un tip de apoyo para complementar su conocimiento. El sistema trata de aumentar la autoestima en el momento que se le facilite y realice correctamente las actividades.

Fig. 6. Pantalla de una actividad de relacionar el concepto con la imagen.

La información de los movimientos del alumno es almacenada en una bitácora para consultas posteriores, es decir, se guarda un seguimiento del alumno y los movimientos

realizados en sus actividades. Al finalizar las actividades, el sistema muestra el nivel con el que terminó y da la opción para borrar su historial. Este nivel está determinado por la cantidad de preguntas que maneja el sistema y los puntos adquiridos con sus respuestas correctas.

Al terminar las actividades del sistema educativo, se tiene la opción de borrar el historial, esta es útil si el alumno desea eliminar toda la información almacenada de sus actividades, la finalidad de esta acción es limpiar la base de datos y regresar al estado original, y de esta forma iniciar nuevamente el curso.

Están registrados en la base de datos los nombres de los todos académicos de dicha materia, se le asignó como password su matrícula única, y su primer apellido.

Las funciones que puede desempeñar un profesor en el Sistema educativo multimedia son las siguientes:

- Ver el contenido temático de la materia.
- Ver la lista de sus alumnos con información de avances.

El administrador puede dar de alta alumnos, profesores, materias y actividades de aprendizaje (ver Fig. 7), el acceso es mediante una clave.

Fig. 7. Pantalla de alta de un curso en la cuenta del administrador en el Sistema educativo.

En la elaboración de las preguntas o actividades a realizar por el alumno se considera el proceso intelectual que se toma en cuenta para cada una de ellas, y dependiendo del tipo de aprendizaje que cada una representa. Dentro de este proceso se le otorgó una ponderación a cada tipo de aprendizaje únicamente como ayuda para establecer el nivel en el que el alumno se encuentra, ya sea, alto, medio o bajo. La ponderación mencionada se observa en la Tabla 1.

Tabla 1. Asignación de puntos por conocimientos.

Conocimiento	Punto
Indispensable	6
Esencial	4
Antecedente	1

Al finalizar las actividades, el sistema muestra el nivel con el que terminó y da la opción para borrar su historial. Este nivel está determinado por la cantidad de preguntas que maneja el sistema y los puntos adquiridos con sus respuestas correctas. Los niveles de conocimientos en metodología de la programación se muestran en la tabla 2.

Tabla 2. Niveles de aprendizaje.

Aprendizaje	Nivel de conocimientos
Nivel bajo	0 a 79 Puntos
Nivel Medio	80 a 149 Puntos
Nivel Avanzado	150 a 166 Puntos

Pruebas del Sistema educativo multimedia Se aplicaron varias pruebas protocolo con el fin de determinar el nivel de conocimientos, habilidades, valores y actitudes con que cuenta el grupo; se revisaron cada una de las pruebas protocolo en base a la escala Likert; además se utilizó los Rangos Numéricos mostrados en la tabla 3, para la interpretación de los resultados. La muestra estaba conformada de 85 alumnos de la FCC de la BUAP de nuevo ingreso, de la generación 2015 y que estaban cursando la materia de Metodología de la Programación en el momento que se realizó la investigación.

Tabla 3. Tamaño Rangos Numéricos para la interpretación de los resultados [4].

No. Pregunta	**0,1,2**	**3,4,5**	**10,11,13,16**	**6,8,9**	**7,14,18**	**12,15,17**	**19,20,I,II,III**
Dimensiones	D1	D2	D3	D4	D5	D6	D7
Media Porcentual de la prueba	3.25	3.95	2.50	3.39	2.38	2.76	2.39
Media porcentual	4.10	4.24	3.35	4.27	3.55	3.58	2.69

Después de aplicar la prueba protocolo 1 y la prueba protocolo 2 a la muestra de alumnos se obtuvieron los resultados mostrados en la tabla 4.

Tabla 4. Resultados de la prueba protocolo 1 de Metodología de la Programación.

Escala	Rango numérico	Valor
Muy bien	4.5 - 5	5
Bien	3.5 - 4.4	4
Regular	2.5 - 3.4	3
Suficiente	1.5 - 2.4	2
Deficiente	1.0 - 1.4	1

En la tabla 4 se muestran los resultados de la prueba protocolo 1 de los alumnos que cursan a materia de Metodología de la Programación y se organizan por dominio. En la Fig. 8 se muestra la gráfica de la media porcentual de cada dimensión de los resultados de la prueba protocolo 1 y protocolo 2.

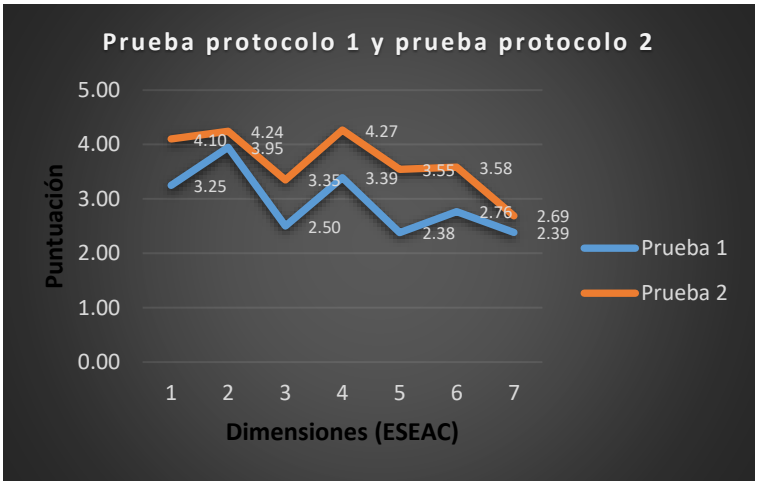

Fig. 8. Media Porcentual de cada dimensión de los resultados de la prueba protocolo 1 y 2.

Enseguida se muestra la descripción de los resultados obtenidos:

El **dominio general del tema** (D1), mejoro y ahora se ubica en el rango de 3.5 - 4.4 con un valor de 4 es Bueno.

El **dominio de lenguajes** (D2), se sigue ubicando en el mismo rango de 3.5 - 4.4 con un valor de 4 teniendo un resultado Bueno.

Calidad de Razonamiento (D3), mejoro y ahora se ubica en el rango de 3.4 - 4.4 con un valor de 4 lo cual nos indica que los alumnos aplican distintas operaciones mentales como son las de analizar, interpretar, sintetizar que permite la solución de problemas y toma de decisiones, siendo en general Bueno.

Errores y su naturaleza (D4), siendo esta área donde el trabajo se establece en un rango de 3.4 - 4.4 lo cual significa que los alumnos no muestran indecisiones e inconsistencias en la ejecución de la respuesta, aunque en determinados aspectos es claramente lógico, ubicándose en un valor de 4, es decir Bueno.

Nivel de Abstracción (D5), mejoro y ahora se encuentra en el rango de 3.5 - 4.4 con un valor de 4 lo cual nos indica que los alumnos mejoraron su proceso mental para realizar una representación mental de la realidad, por lo que no cometen errores de importancia al realizar dichas representaciones siendo en general Bueno.

Conciencia Cognitiva (D6), mejoro y ahora se encuentra en el rango de 3.5 - 4.4 con un valor de 4 lo cual nos indica que los alumnos mejoraron sus operaciones mentales y cognitivas como son de percepción, observación, razonamiento y aprendizaje.

Nivel de motivación y ansiedad (D7), se establece en un rango de 2.5 - 3.4 lo cual significa que los alumnos disminuyeron algunas indecisiones e inconsistencias en

el manejo de sus emociones para realizar actividades o tareas, sin embargo, seguimos requiriendo auto-motivación y mejorar sus relaciones interpersonales ubicándose en un valor de 3, es decir Bueno.

La competencia que más se desarrolló fue "Resolver problemas de manera autónoma" con un 82% de desarrollo, de 65% a un 82% y la segunda fue "Organización del pensamiento" con un 67%, logrando alcanzar de un 50% a un 67% esto conlleva que el uso de las estrategias propuestas propicia un aprendizaje más significativo en el tema de Metodología de la Programación.

4. Desarrollo y pruebas del sistema educativo multimedia para el apoyo del aprendizaje autónomo de metodología de la programación

Una de las principales contribuciones del Sistema educativo multimedia de corte autónomo para el aprendizaje de metodología de la programación (SEMMP) es la adaptación de distintos perfiles de usuarios al aplicar la actividad de aprendizaje.

El SEMMP promueve en los estudiantes un aprendizaje contextual, experimental, participativo y de autoaprendizaje y a apoyar a los docentes, ya que motiva las competencias tecno-pedagógicas al incluir el uso de las TIC en su práctica docente.

Con base a la información obtenida durante la interacción del alumno con el sistema, es importante observar que se fomentó el trabajo en equipo, la participación y el aprendizaje autónomo.

Una de las principales perspectivas de este trabajo es elaborar Sistemas educativos adaptativos de otras materias y otros niveles educativos: educación básica y media superior, integrando nuevas tecnologías como son los agentes inteligentes, simulación que logren tareas más específicas y puedan establecer relaciones con otros aspectos del aprendizaje (abstracción, detección de errores, etc.).

Referencias

1. Crispín, M.L. et al.: Aprendizaje Autónomo. Orientación para la docencia. Universidad Iberoamericana, México (2011)
2. Vera, E.: Estrategias de aprendizaje autónomo para disminuir los índices de reprobación en la materia de metodología de la programación en la FCC de la BUAP en un sistema de e-learning. Tesis Doctorado, México, CECE (2017)
3. Bernad, J.A.: Modelo Cognitivo de Evaluación Educativa. Escala de Estrategias de aprendizaje Contextualizado (ESEAC), Narcea, S.A. de Ediciones Madrid (2007)
4. García, A., Gómez, E.: Planeación y organización de secuencias e intervenciones didácticas en el aula (DIPFEC). México, 1–20 (2012)
5. Ganem, P., Gutiérrez, A.: Elaboración de exámenes con reactivos constructivistas. Limusa noriega (2013)
6. Piaget, J.: Piaget's Theory. En P. H. Mussen (Ed.): Carmichael's Manual of Child Psychology (Vol. 1), New York Wiley (1970)

7. Vygotsky, L. S.: Mind in society: The development of higher psychological processes. Cambridge, MA, Harvard University Press (1978)
8. Ordóñez C.: Pensar pedagógicamente desde el constructivismo. De las concepciones a las prácticas pedagógicas, Revista de Estudios Sociales no. 19, diciembre de 2004, 7–12 (2004)
9. Amaya, G.: Aprendizaje Autónomo y Competencias. Bogotá (2008)
10. Lara, A.: Aprendizaje Autónomo para estudiantes Universitarios. Colombia: Editorial Universidad de Caldas, Artes y Humanidades (2008)
11. PC World en Español-Digital Too: Factory Pyme. The Standard IT-TV y el portal The Standard IT, 1–26 (1993)
12. González, A., Castro, E., Bañuelos, D.: Trayectorias escolares. El perfil de ingreso de los estudiantes de Ciencias Químicas: un primer abordaje para contrastación ulterior con otras disciplinas. Distrito Federal, México (2011)
13. Cáceres, A.L.: Perfil del estudiante sobresaliente del bachillerato de la UNAM. Características generales de la población, México, UNAM (1992)
14. Quesada, C, Sánchez, J.: Calificación y Diagnóstico del Aprendizaje por Computadora. México, Noriega Editores (LIMUSA) (1996)

ISSN 1870-4069

Genex+, a Semantic-based Automatic Extractor of Examples Applied to Bilingual Terms

Jorge Lázaro[1], Juan-Manuel Torres-Moreno[2,3], Gerardo Sierra[4], Teresa Cabré[5], Andrés Torres Rivera[2,5]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de Filosofía y Letras, Puebla, Mexico

[2] Université d'Avignon et des Pays de Vaucluse, Laboratoire Informatique d'Avignon, Avignon, France

[3] Ecole Polytechnique de Montréal, Département de Génie Informatique et Génie Logiciel, Quebec, Canada

[4] Universidad Nacional Autónoma de México, Grupo de Ingeniería Lingüística, Instituto de Ingeniería, Mexico

[5] Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, Barcelona, Spain

jorge.lazaro@correo.buap.mx, juan-manuel.torres@univ-avignon.fr, gsierram@iingen.unam.mx, teresa.cabre@upf.edu, andres.torres01@estudiant.upf.edu

Abstract. In this paper we present *Genex+* (Genex plus), an improved implementation of the Genex system (Générateur d'Exemples) [17], by using the semantic closeness between certain fragments of texts. This process was based on the combination of the successive extraction of concordances and collocations with the determination of the semantic closeness by the Mutual Information method [24] and Cosine calculus [3,22]. Results have proven that a good example is almost always associated to the definition of the word to which it is making reference, and that it can be extracted automatically by the consecutive restriction of semantic fields applied to fragments of general language corpora. This has the advantage of presenting a conceptual reformulation of what is said in the definition by using highly informative textual fragments, but with a less formal register. For this second version, we implemented changes required for a bilingual entry.

Keywords. Terminology exemplification, computational lexicography, lexicometrical density, semantic saturation.

1 Background

1.1 Exemplification in Bilingual Dictionaries

When bilingual dictionaries are being planned, the design of correct examples has been a constant concern because of the pedagogical and supportive functions they perform

[1,6,8,10,18,19,20,21]. Nevertheless, the selection of good examples [2] from corpora represents a difficult task that needs techniques and tools to overcome the selection of candidates that could potentially be good examples.

Regarding examples in foreign language (L2) dictionaries, we have based this article on Philippe Humble's proposals[12]. In his study, Humble emphasizes that an example will not be unique for all users, rather that the words surrounding the term to be exemplified will depend on the training level of the users, which will be divided into two groups. The first group includes beginning learners who need the example to comprise very frequent words with very frequent meanings. While the second group includes advanced learners who need the example to comprise less frequent words or less frequent meanings of frequent words. Furthermore, the author mentions that the first group would use *made-up* examples, while the second group would use authentic examples (those retrieved from corpora). Similarly, Humble mentions that these authentic examples can be used to illustrate the use of intermediate frequency words.

However, even with these accurate notes, we must consider that there is no consensus on the exact source of examples. This is, generally the source that tends to be general corpora, this is not always a decisive criterion. This is important, because given that the exact source of an example is unknown, then there is no way of knowing if the speaker of the phrase to be used as an example or the text from where it is extracted differs from the definition of the term. By this, we mean that definitions and examples are drawn from different sources. One might think that the specialized dictionaries would be different because they need more detail to convey specialized concepts, however, both practices are very similar and the origin of examples and definitions is definitely not the same[17]. Furthermore, there are studies which mention that terminology or specialized units should be treated equally when referring to the example associated with them[25], as if examples for a word of general use and examples for a term followed the same process of association both with the term to which they refer, and with the reality they are a part of, in order to reflect the use of the term in the language.

In this sense, neither studies relating to examples in terminology (which are very few), nor studies relating to examples in lexicography, mention the source from which these fragments should be extracted, even when all the features they should have are mentioned. However, a quick scan shows that, because of the information or their function[11], their source is most relatable to a general language corpus.

In the literature devoted to the study of the subject at hand, a point where all perspectives converge has not been found and an exact point where the example could be ideally located and cannot be indicated. Previous research[17] has concluded that the example does not seem to be there *a priori*. To wit, the selected fragments are always related to a set of functions of complementary nature that must comply with regards to the discourse, and that adopts a better understanding of their reference, whether is a term or a word. Therefore, it is difficult to speak of a *location* in the strict sense of the word, because selected examples for dictionaries are most likely to be adaptations of textual fragments found in reference corpora or are creations based on the frequency of use of the words surrounding the term to exemplify.

However, without being there *a priori*, the example can be *identified* – and this is the key word – thanks to the conceptual complementarity needs, it covers regarding the term

and how a receiver or person consulting a dictionary apprehends it. "Complementarity needs" refer to the set of communication situations where the term is used and which readers need to associate with their every day lexicon. This is, to complement the meaning of a term (either very known or not) means locating the word in linguistic contexts where other surrounding words imply semantic relations which may lead the apprehension process of the meaning of the word to exemplify. Therefore, we consider that an example is not only a context where a term or word can be found, rather it is a context where the term or word is enunciative. In this sense, it is therefore basic to consider that the example has certain formal and semantic features that allow it to fulfill its role of *revealing* these semantic relations. The functions it performs, regarding the clarification of the meaning of a word, are based on the functions that each of the surrounding words in the exemplifying fragment establishes.

From the literature mentioned above, it is clear that current work on example identification and extraction from corpora makes use of the criteria we have mentioned, but without systematizing them, let alone naming them, meaning, it is done intuitively. A lexicographist or terminologist looks at how functional and potentially adaptable the documented fragments are (giving them their witness category) in order to be associated with a word, but does not consider the associations or relations between the word, its definition, and the missing information so that any reader can associate the word with a specific meaning. In this proposal we will focus only on examples associated to terms. We will highlight the identification and extraction of examples in specialized dictionaries, and we will consider that current examples have been designed as an adaptation of something existing – or documented –, that is, that they necessarily come from a verifiable source, such as corpora. We will also acknowledge that we have designed them as a pragmatic necessity, but mostly as a communicative necessity, because selecting the ideal textual fragment to exemplify a term must be cohesive internally (all its elements must be related), as well as externally (these elements must also relate to the definition of the exemplified term).

Therefore, based on the identification of examples associated to terms and their communication implications, we will be working specifically based on the Communicative Theory of Terminology [4], to try to explain the behavior of examples in terminology dictionaries. First, we will review the theoretical principals that will allow us to outline the notion of examples in terminology, and then we will show how we apply this proposal in the design of a tool that automatically extracts examples based on their semantic and syntactic features. Furthermore, we will show that a translator module, which allows users to generate examples in a source language based on terminological equivalents in a target language, has been added to this extractor.

1.2 Automatic Extraction of Examples

Many papers on the selection and extraction of examples for dictionaries show that human monitoring especially for linguistic issues (the most common problem being cohesion and coherence in phrases) is still needed; although human influence is trying to be kept to the minimum. For the development of the tool we are now presenting, we have relied on one of the most outstanding works in this field, GDEX or *Good Example*

Extraction[13]. This system tries to be a model of semi-automatic extraction of examples from large textual corpora. In former models, examples mainly followed the criteria of different lexicography manuals and treaties on dictionaries, which are summarized in Atkins and Rundell's work[2]. With this classic methodology as working principal, it is not unusual that some researchers designed tools to support the semi-automatic extraction of examples, consisting in obtaining concordances for collocations that were stored and analyzed by human agents, one-by-one, in order to find a good example. GDEX's method, however, retrieves classified concordances and presents them in order of importance. In order to classify them this way, it assigns weights obtained from a list that students have previously rated (based on their linguistic knowledge) which is then evaluated in a way that the system selects the same options as students, creating rules to select the best collocation[1].

The results suggest an interesting attempt, but the contextualization of the word to exemplify remained a difficult problem to overcome. In other words, GDEX extracts relevant fragments, but their relevance remains as a classification problem given that it cannot be completely done automatically because ambiguities often need to be solved. Research on other languages has obtained similar results which are satisfactory, but not always appropriate. Some cases worth mentioning are Slovenian[14] and Swedish[23]. In the case of German[7], the authors resorted to works previously done for GDEX and adapted the notions of *frequency*, *length*, and *instance of the word "matrix"*, which they summarized in two criteria: readability and complexity[2]. However, it is worth mentioning that this system works acceptably when used for lexicographical purposes, but results are not as satisfactory when used for terms or specialized words.

These approaches show that a certain type of pattern can be found, and that, by the combination of *concordances + collocations* a series of reductions could be used to determine the instance of certain lexical elements that complement the information of a word. These are complex concordances that show specific contexts, of which we will speak later.

1.3 Genex: Générateur d'Exemples

The Genex system works on the basis of an association measure called *lexicometrical density* [17]. This measure is a direct product of the notion of *semantic saturation*[3]. Semantic saturation is a theoretical framework which basically states that ideally a person knows a term in all its instances, its definitions, its contexts, and all the conceptual variants that it may have within a specialized discourse. If we include the above within the Communicative Theory of Terminology[4], which we are currently working with, this would be the same as saying that a person knows enough to cover all the sides of

[1] "Once the features have been identified, the question arises: how should they be weighted? Which features are most important, and by how much? With this in mind, we asked two students to select good examples for 1000 collocations. We then used those "known good" examples to set the weights by automatically finding the combination of weights that would give the "known good" examples the highest average rank. The first two features, sentence length and word frequencies, were given greatest weight"[13]

[2] `http://www.natcorp.ox.ac.uk/`

[3] To observe the specific use of this measure in exemplification refer directly to (Lázaro, 2015)

the polyhedron which represents the concept of a term, that is, the person would ideally fulfill the *Polyhedral principle*[5]. To measure *semantic saturation*, it is essential to acknowledge that the example is closely related to the definition of the term, and that the term's elements should appear ideally in the former, but not in a strict word order or under the same lexical rules. Meaning, there may be variation. Precisely this relation measures the *lexicometrical density*. The items that are measured in the example are those words present in the definition excluding defining verbs or function words. The term, in this theory, is an essential element.

The instance in the example of words from the definition addresses two main issues. On the one hand, the inclusion of the term in the example gives the fragment a conceptual nucleus around the elements of the definition revolve, which will ultimately be satellites and limits to the significance of that term. Seen in this way, the combination of the words of a definition in a new non-defining structure allows the term's activation context[15] to continue operating. A term's activation context is a theoretical concept that allows seeing any given word within a context where it acquires a specialized value. This is, a context where it is already being used as a term. This kind of context is, thus, a fragment which makes a given word a member of a particular semantic field. These fragments, however, are usually simple collocations of the given word. From this perspective, applying another filter is needed in order to find complex collocations.

Not every activation context is a sentence. To give cohesion to a new fragment which could potentially work as an example, only those that contain a verb semantically close to term will be chosen. This verb can exist *a priori* in the definition (a verb other than definitional) or may be the result of a search made within the working corpus through the *Mutual Information measure*[24]. This filter allows finding complex collocations of the terms to exemplify.

As it can be seen, the first part of the *lexicometrical density* calculates how much information a textual fragment has and how much semantic proximity it has with another fragment, so that the former can complement the latter at the conceptual level. In our case, we consider a definition to be a vector formed by its conforming elements, all the words. The second vector, the closest one, would be a potential example. This association is made through the cosine measure[3,22]. The cosine determines that a sentence is semantically close by the amount of information it contains, but does not take into account its length. Thus it is not difficult to infer that many of the fragments extracted with this methodology were too long, even more than the definitions to which they were associated. Therefore, we decided to choose only those fragments with high scores, but with fewer tokens. Having determined this restriction, we can consider that *lexicometrical density* is the product of the cosine value of a sentence (any with words from the definition) regarding another one (the same definition) by the inverse of the logarithm of its length. That is, the second part of the algorithm associated with this formula extracts the shorter sentences but which contain as much information as possible. Graphically, Formula 1 was used to determine the aforementioned:

$$score_i = \cos(sentence_i, q) \times \frac{1}{\log |sentence_i|}. \tag{1}$$

The Genex system uses four basic resources:

1. A general language dictionary.
2. A set of corpora.
3. A morphological tagger (TreeTagger).
4. A program to calculate Mutual Information.

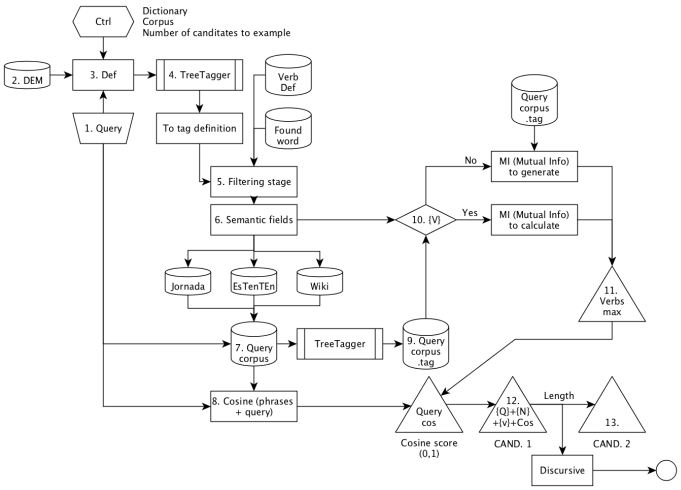

Fig. 1. Genex's Architecture.

As it can be seen, the program basically operates in the following way:

- The user inserts a term (1).
- The program asks to choose a meaning (2 and 3) from a pre-loaded dictionary.
- The program creates the nominal context subcorpus from lexical elements.
- the definition (4-9) from a general corpus.
- The program creates the verbal context with the verbal elements of the definition that do not belong to an exclusion list consisting of defining verbs in Spanish (10).
- If there are no verbs to continue the search, the system determines which verb is closer to term within a subcorpus including all the phrases containing one or more elements of the nominal context. In order to make this determination, it uses the measure Mutual Information (10 and 11).
- Once the nominal and verbal contexts are determined, it chooses those phrases that meet the restrictions: they unfailingly include the term, they contain one or more words of the nominal context, and they contain at least one element of the verbal context.
- It evaluates the chosen phrases using the *cosine measure*[4] to determine which one is closer to the definition.

[4] "...examples are ordered in respect to their goodness with the help of some soft criteria which are listed in the order of their importance: a) including words should be among the 17,000 most frequent words of our balanced corpus; b) including words should be no longer than 15 characters; c) finally, the keyword should be within the matrix clause."

– Among these, it chooses those that, in addition to meeting all the criteria, have fewer tokens.

All these criteria under which Genex works, have allowed us to obtain a methodology to extract those complex concordances of specific contexts of which we spoke earlier. This type of concordances, or phrases as they are called throughout the study, will be known as a *candidate for example*. This list of candidates fairly reliable from the lexical-semantic perspective, from which lexicographers or terminologists can choose an example suitable for the type of dictionary they are creating.

2 Genex+ and its operation

In this new stage, Genex's operation was modified to be used with several languages. Basically, a term translator was added so that equivalences of a target language could be found in a source language. For the translator to work, dictionaries in both target and source languages were needed. Finally, once the system was able to determine which and how many equivalents were relevant for each term, it proceeded to work as usual. This optimized language-independent system (it can work with several languages by only changing the dictionaries) is called *Genex+* (*Genex plus*). To test the effectiveness of Genex+, an experiment was performed with terms from the field area of oncology in two languages, French and Spanish.

2.1 A Term Translator Module

Machine translation (MT) is one of the central issues of Natural Language Processing (NLP). Although several MT issues remain unresolved, research in NLP has lead to the development of MT systems with an increasingly better performance.

A few years ago, MT systems used rules to translate from one language to another. This approach has several limitations, such as the difficulty to formulate rules for every language, or the need for experts who can write and code such rules in appropriate programs. Currently, statistical machine translation (SMT) has evolved with the aid of algorithms which use the large body of available corpora (both aligned and unaligned) to calculate the translation probability of phrases, fragments of phrases, or words.

The challenges of MT also extend to the semantic level, which remains the most difficult to overcome. For example, with the sentence *"Je vais acheter un canapé"*, how do we know if the speaker is hungry or requires furniture? The algorithm should be able to decide and propose a translation for *"canapé"* in Spanish to choose between *"sandwich"* or *"sofá"* (sandwich or couch). Or in *"j'ai vu la dame avec les jumelles",* how do we know if the lady was seen along with twin sisters or if the speaker used binoculars to look at the lady? In the case of English, once again choosing between *"I saw the woman with the twins"* or *"I saw the woman through the binoculars"* is an issue that has not been completely solved with current systems. We consider that the automatic extraction of examples may help to clarify the meaning of words – and therefore improve interpretation – if the extracted fragments come from the definitions of the exemplified word. That is, a possible disambiguation method would rely on the

direct observation of the candidates for examples of the given lexical item (canapé, lady, furniture, and so on).

In the framework of the Genex example generator, we wanted to enable the possibility to enter a term in a language L1 and obtain candidates to exemplify in another language L2. In the occurrence, L1 = English, French and L2 = {Spanish}. As might be expected, this function should imply the least number of changes on the system resources (parser, example ranking algorithm, etc). The simple solution that was used was a word-by-word translation for the entry terms. Figure 2 shows the simplified architecture of the system. Genex's core (algorithm + Spanish corpus + TreeTagger) remains intact, and the translation module is only an interface in the system input. At the moment, only one-word terms are used to generate examples, therefore MT syntax and semantics problems are irrelevant to our approach. Bilingual dictionaries from another language to Spanish let the system generate the examples using the methodology described below. Currently, a multiterm module at the input is being developed.

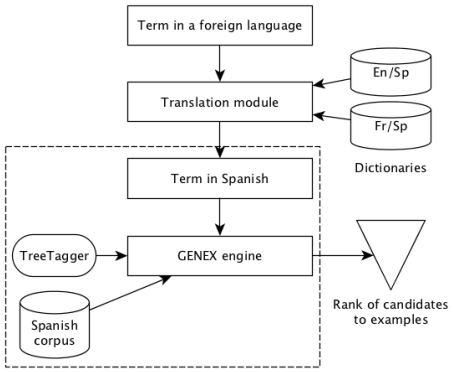

Fig. 2. Simplified Architecture of Genex+.

2.2 Dictionary Preprocessing

Lists of equivalences terms for English-Spanish and French-Spanish were required to carry out the corresponding tests with Genex+. These lists should proceed from an official source, thus the *Inter-Active Terminology for Europe* (IATE)[5] database was selected. This database contains 8.4 million terms and only includes the 24 official European Union languages, therefore Catalan could not be included.

IATE offers a distribution of their database with approximately 1.4 million entries and around 8 million terms in total. This database is downloaded as a TBX compressed file that can be consulted using a Java application which generates a subcorpus with the desired parameters introduced by the user. In this way, two working subcorpora were obtained, one for English-Spanish and one for French-Spanish, with equivalences

[5] http://iate.europa.eu/

shared by L1 and L2; this is, not all terms in Spanish have the same equivalences in other languages.

Once the working subcorpora were obtained, they were formatted into equivalence lists enabled for Genex+. IATExtract generates lists in TBX format[6] that are XML-based and allow information retrieval with software applications developed for this purpose or using a XML parser.

Because we required generating lists and not querying the corpus, a parser was developed in Python by implementing the *xml.sax*[7] library, to generate files with the format required by Genex+; only the tags that were necessary were kept and the rest were ignored. The work flow of the script in pseudo code is described below:

```
Read the files to be processed
Generate a new file
Find term code tags
Find language code tags
For each term tag
    If L1 and L2 share a term tag
        Group L1 and L2 terms by term code
        Write the data on the new file
        Apply the necessary format
Close used files
```

A first attempt was performed with the TERMCAT terminological dictionary, but given that the definitions were in Catalan, it could not be used by Genex+. For this reason, the terminological dictionary was obtained from the Spanish Association Against Cancer[8] and it was processed through the same procedure that had been applied to the IATE database. The dictionary's source code was downloaded and because it had been formatted as text in HTML, its processing was simpler: the HTML tags formatting the entries and definitions were read and the information was structured to match the format required for Genex+ afterwards. Once these elements were incorporated, the corresponding tests were carried out. The appearance of a dictionary edited with this methodology can be seen in figure 3.

2.3 Extraction of Examples Through Equivalent Terms

Once the new dictionaries and the lists of equivalences were obtained, we proceeded to generate examples in the target language (Spanish) from equivalences in the source language (French). The appearance of Genex+'s main menu can be seen in Figure 4.

Below are the results of five oncology terms in French, whose equivalents have been used to create a list of example candidates. There are only 2 cases of each term.Complete results for each one of them are shown in the Appendix[9]:

[6] http://iso.org/iso/catalogue_detail.htm?csnumber=4579

[7] https://docs.python.org/3.4/library/xml.sax.html

[8] https://www.aecc.es

[9] The experiment was carried out with 5 terms, resulting in a total of 25 example candidates. 44 people evaluated the 25 candidates, which yielded a total sample of 1100 variables.

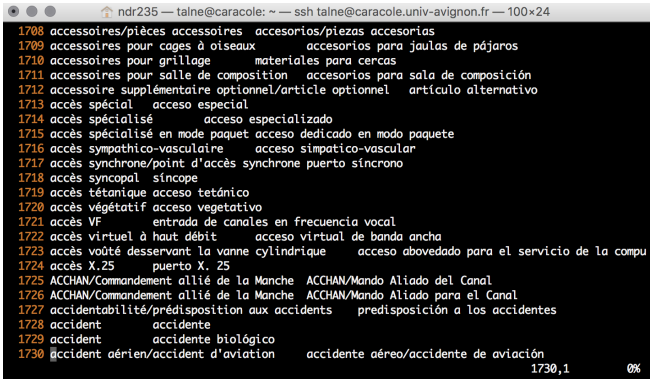

Fig. 3. Dictionary of Equivalent Terms.

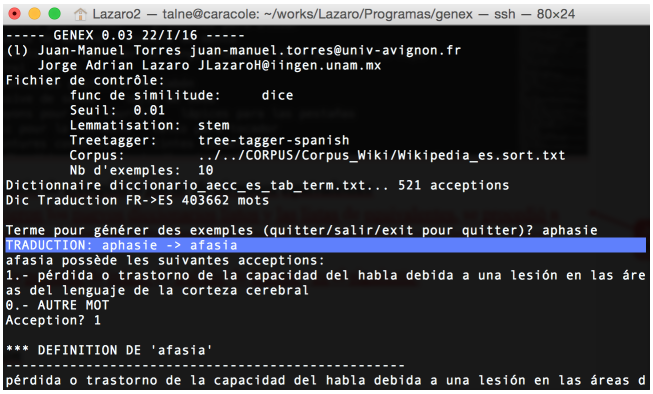

Fig. 4. Genex's Main Menu.

aphasie > afasia

1. Debido a esto, quedó traumatizada y durante dos años padeció afasia (pérdida del habla). 0.0889
2. Se distingue de una afasia motora en que no es un trastorno del lenguaje, sino del habla; es decir, el paciente manifiesta dificultades asociadas con la articulación de fonemas. 0.0815

carcinome > carcinoma

1. Estos son tumores caracterizados por células epiteliales grandes con citoplasma claro y abundante y con frecuencia se ven asociados con endometriosis o con un carcinoma endometrioide del ovario y con gran semejanza al carcinoma de células claras del endometrio. 0.0774
2. El carcinoma anaplásico de "células en avena" es el tipo histológico que con mayor frecuencia se asocia a la producción ectópica de hormonas y a síndromes paraneoplásicos. 0.0657

chimiothérapie > quimioterapia

1. Y es que los tratamientos a base de quimioterapia, de la que existen alrededor de 50 productos diferentes, si bien combaten las células cancerígenas, también destruyen las sanas. 0.0468
2. "En los casos en los que los pacientes sufren de enfermedades como el cáncer, tumores, melanoma, diabetes, sida, artritis reumatoide, alzheimer, esclerosis múltiple o lupus eritematoso severo pueden ser tratados con radiación o quimioterapia para destruir las células anormales. 0.0421

leucémie > leucemia

1. La leucemia linfoblástica precursora aguda de células B es un tipo de leucemia linfoide aguda que afecta en particular los precursores de los linfocitos B que están localizados en la médula ósea. 0.0882
2. Sin embargo, en algunos tipos de leucemias también pueden afectarse cualquiera de los precursores de las diferentes líneas celulares de la médula ósea, como los precursores mieloides, monocíticos, eritroides o megacariocíticos. 0.0641

tumeur > tumor

1. A veces el IGF -2 es producido en exceso por tumores de células islotes, causando hipoglucemia. 0.0815
2. Los tumores derivados de células del intestino posterior (carcinoide rectal) raramente producen un exceso de 5 -HIAA. 0.0774

2.4 Evaluation

The evaluation was carried out with the help of human agents. In total 44 people tested the relevance of the results provided by Genex+, of which 31 were women and 13 were men. Of all respondents only 6 did not hold a bachelor's degree, so they were considered a set of educated speakers. 95% of them were Spanish native speakers. Their average age was 31.24 years old.

For each term, 5 examples were extracted and each candidate was evaluated by the 44 evaluators. A total of 1,100 variables or answers were obtained to weight the relevance of the examples associated to the term. *The Pearson's correlation coefficient*, a measure of the linear correlation between two quantitative variables[9], was used to calculate the effectiveness of the system. This measure provided the degree of relation among the two variables: relevance of the examples produced by the system according to the criteria it uses and preference of speakers regarding the information these examples provide about the definition.

Pearson's correlation coefficient is represented by $\rho_{X,Y}$ and is calculated through Formula 2:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \tag{2}$$

where XY is the covariance of (X,Y), σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y.

Cohen's Kappa correlation measures the agreement between two evaluators, or two evaluation methods. It is represented by κ and calculated using Formula 3, where P_o is the observed agreement among raters and P_e is the hypothetical probability of chance agreement:

$$\kappa \equiv \frac{P_o - P_e}{1 - p_e} = 1 - \frac{1 - P_o}{1 - p_e}. \tag{3}$$

The data obtained is to be interpreted as the existent correlation between the system choices and what speakers prefer. These results range between 1 and -1. A value close to 1 means that the system has succeeded in the selection of a candidate for example, the more distant it gets, that is, when it gets closer to -1 it means the system failed at ordering as humans do. A value of 0 implies certain similarity between the human and system preference/aversion towards certain results.

The following table shows the results obtained by Genex+. The example's number corresponds to the order shown on the previous section, 1 being what the system considers an ideal example and 5 the least likely example according to their *lexicometrical density*. As it can be seen in Table 1, the Pearson column shows the extent to which the human evaluators agreed with this arrangement.

3 Discussion and Conclusions

As it can be seen, the system had a good performance overall, as both human evaluators and the system results mostly agree. There was an agreement above 0.3 almost on every instance, which indicates a trend towards total agreement. Furthermore, there are cases with agreement quite close to 1. For instance, according to this evaluation, *"chimiothérapie"* and *"leukemia"* obtained values as high as 0.7.

As in all automatic systems, we find complex cases, such as the example candidates for the term *"tumeur"* were evaluated almost inversely, in other words, closer to -1. Regarding this particular instance, it is worth mentioning that there is no aversion towards the candidates proposed by Genex+, but rather to their *lexicometrical density* ranking. Apparently human agents had a preference for the length of the example over its density. This is to be expected if we take into account that most of the times the explanatory function of the example or its pragmatic extension function are hampered by the brevity of a sentence. As can be seen, the most accepted example candidates (3 and 5) for this term contain words associated to the term, as well as specifications of the functions or affectations of *"tumor"*. This information can often be curtailed if we consider only the definition to which an example is associated and not its common use. Although the tool associates this fragment to the specialized definition in an acceptable way, and brings it to a less formal register, the fact is that there are also information specifiers or magnifiers that do not occur at word level (a word that determines another in a specific context, which is how Genex+ works) but at discursive level, where a whole sentence specifies a single word of another sentence; all of which complicates the analysis. This, as can be inferred, complicates the automatic processing of examples, because we would have

Table 1. Results obtained by Genex+.

Example	Human Response	Pearson	Kappa
aphasie >afasia			
1	9.38%	0.4	0.0
2	37.50%		
3	31.25%		
4	15.63%		
5	3.13%		
carcinome >carcinoma			
1	43.75%	0.3	0.25
2	12.50%		
3	15.63%		
4	0.00%		
5	25.00%		
chimiothérapie >quimioterapia			
1	34.38%	0.7	0.25
2	28.13%		
3	12.50%		
4	6.25%		
5	15.63%		
leucemie >leucemia			
1	28.13%	0.7	0.25
2	6.25%		
3	50.00%		
4	6.25%		
5	6.25%		
tumeur >tumor			
1	0.00%	-0.9	-0.25
2	3.13%		
3	40.63%		
4	6.25%		
5	46.88%		
Average		**0.24**	**0.1**

to think that sometimes the meaning of a term and its exemplification is not only in the phrase where it can be inserted, rather this exemplifying information is shared along larger structures such as subordinate clauses. Even with this, there are few cases that yield such results, so we can conclude that overall the Genex+ system works correctly and humans consider it an acceptable tool for selecting candidates as examples.

Finally, the results suggest that the best examples in a bilingual dictionary would be associated with the definition of the target language and not to that of the mother tongue[16], because the linguistic system is different. However, examples in both languages would complement each other to attest the way in which reality is apprehended

by the language that is required to know, as it has already been shown by Humble[12]. To achieve this, we could reverse the dictionaries and corpora to obtain a bidirectional tool, which is completely possible if it is taken into account that Genex+ is a language-independent system. This means that it would help the learner to discern those contexts where an equivalent can be used, and when this is not possible because the contextual usage restrictions have been automatically applied previously by the program.

Appendix: Complete list of results for each term

aphasie > afasia

1. Debido a esto, quedó traumatizada y durante dos años padeció afasia (pérdida del habla). 0.0889
2. Se distingue de una afasia motora en que no es un trastorno del lenguaje, sino del habla; es decir, el paciente manifiesta dificultades asociadas con la articulación de fonemas. 0.0815
3. Las disritmias son un trastorno de emisión, al igual que las dislasias y disartrias (trastornos de articulación por anomalías en los órganos articulatorios que pueden solucionarse con logopedia), disfasias y afasias (deterioro al adquirir el habla y pérdida de la misma) y disfonías y afonías (anomalías en la producción por afecciones laríngeas). 0.0653
4. El descubrimiento por este último de que las lesiones localizadas en una región del lóbulo frontal izquierdo (llamado después "área de de Broca") podrían entrañar una afasia (una incapacidad de hablar) infligieron un duro golpe a la doctrina holística. 0.0502
5. La afasia de Wernicke no sólo afecta a la comprensión del habla. 0.0491

carcinome > carcinoma

1. Estos son tumores caracterizados por células epiteliales grandes con citoplasma claro y abundante y con frecuencia se ven asociados con endometriosis o con un carcinoma endometrioide del ovario y con gran semejanza al carcinoma de células claras del endometrio. 0.0774
2. El carcinoma anaplásico de "células en avena" es el tipo histológico que con mayor frecuencia se asocia a la producción ectópica de hormonas y a síndromes paraneoplásicos. 0.0657
3. Los factores de riesgo de cáncer cervical están relacionados con características tanto del virus como del huésped, e incluyen: Varios tipos de VPH, particularmente el tipo 16, han sido hallados asociados con carcinoma orofaríngeo de células escamosas, una forma de cáncer de cabeza y cuello (en inglés). 0.0513
4. Se ha sugerido que un SNP extraño (rs11614913) que está superpuesto a la hsa-mir-196a2 está asociado al carcinoma pulmonar de las células pequeñas. 0.0461
5. Los tumores malignos, por su parte, son del tipo adenocarcinoma quístico formando un 40% de todos los carcinomas malignos de ovario, por lo general se ven en mujeres avanzadas de edad, frecuentemente asociados a casos familiares y un 66% de los casos de tumores malignos de ovario son bilaterales. 0.0410

chimiothérapie > quimioterapia

1. Y es que los tratamientos a base de quimioterapia, de la que existen alrededor de 50 productos diferentes, si bien combaten las células cancerígenas, también destruyen las sanas. 0.0468
2. "En los casos en los que los pacientes sufren de enfermedades como el cáncer, tumores, melanoma, diabetes, sida, artritis reumatoide, alzheimer, esclerosis múltiple o lupus eritematoso severo pueden ser tratados con radiación o quimioterapia para destruir las células anormales. 0.0421
3. Con los conocimientos acerca del efecto en conjunto de los tres genes se podráinvestigar de manera más precisa cómo adquieren las células cancerígenas las características con las que logran sobrevivir en el cuerpo, y probablemente promover la creación de medicamentos para cada tipo de cáncer, lo que disminuiría el uso de la quimioterapia y la radiación, que destruyen también el tejido normal. 0.0355
4. Y también los medicamentos y remedios necesarios se aplican bajo diferentes esquemas: los productos que destruyen los vasos sanguíneos se aplican por largos periodos, mientras que la quimioterapia se aplica por ciclos. 0.0188
5. Discos magnéticos ultrafinos de un micrón de diámetro y unos 60 nanómetros de espesor pueden destruir células cancerígenas, sin los efectos secundarios de las quimioterapias, indicó un estudio publicado el domingo por la revista científica Nature Material. 0.0148

leucémie > leucemia

1. La leucemia linfoblástica precursora aguda de células B es un tipo de leucemia linfoide aguda que afecta en particular los precursores de los linfocitos B que están localizados en la médula ósea. 0.0882
2. Sin embargo, en algunos tipos de leucemias también pueden afectarse cualquiera de los precursores de las diferentes líneas celulares de la médula ósea, como los precursores mieloides, monocíticos, eritroides o megacariocíticos. 0.0641
3. Madrigal considera que la leucemia -cáncer de sangre- no afecta a células secundarias sino a las células madres hematopoyéticas que son las que funcionan mal en los pacientes enfermos. 0.0641
4. En la leucemia mieloide crónica la translocación genética de los cromosomas 9 con 22 da lugar a una tirosincinasa que ejerce una importante acción en los mecanismos de adhesión, apoptosis y proliferación de las células mieloides afectadas. 0.0392
5. Los precursores de linfocitos B afectados tienen una serie de receptores en su membrana, lo cual permite identificar el tipo de leucemia como una leucemia linfoblástica precursora aguda de células B. Otras enfermedades: diagnóstico de leucemia aguda linfoblástica hace 5 años recibió quimioterapia se desconoce número de ciclos y medicamentos utilizados. 0.0390

tumeur > tumor

1. A veces el IGF -2 es producido en exceso por tumores de células islotes, causando hipoglucemia. 0.0815

2. Los tumores derivados de células del intestino posterior (carcinoide rectal) raramente producen un exceso de 5 -HIAA. 0.0774
3. Se caracteriza por el crecimiento de tumores benignos que puede desarrollarse a partir del tejido de un solo linfonodo o a partir de múltiples sitios simultáneamente El crecimiento de los linfonodos radica en la hiperproliferación de ciertas células B que con frecuencia son productoras de múltiples citoquinas. 0.0608
4. Se inyecta calcio en el páncreas, lo que causa que las células beta liberen insulina; si hay un tumor que causa un exceso de células beta, la insulina se producirá en demasía y el azúcar caerá demasiado y abruptamente. 0.0591
5. Warburg hipotetizó que el cáncer, el crecimiento maligno y el crecimiento de los tumores son causados por el hecho de que las células tumorales generan energía (producen ATP) principalmente por medio de una degradación no oxidativa de la glucosa (un proceso llamado glicólisis anaeróbica; al contrario de lo que ocurre con las células saludables, las cuales generan energía principalmente a partir de la degradación oxidativa del piruvato. 0.0580

References

1. Alvar Ezquerra, M.: Diccionario y gramática. LEA: Lingüística española actual 4(2), 151–212 (1982)
2. Atkins, B.T., Rundell, M.: The Oxford Guide to Practical Lexicography. Oxfor University Press, Oxford (2008)
3. Baeza-Yates, R., Ribeiro-Neto, B.: Moder Information Retrieval. ACM Press Addison-Wesley, New York (1999)
4. Cabré, T.: La terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona (1999)
5. Cabré, T.: El principio de la poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en terminología(i). Ibérica pp. 9–36 (2008)
6. Cowie, A.P.: The Language of Examples in English Learners' Dictionaries, pp. 55–65. No. 14, University of Exeter, Exeter (1989)
7. Didakowski, J., Lemnitzer, L., Geyken, A.: Automatic example sentence extraction for a contemporary german dictionary. In: Eurales 2012. pp. 343–349 (2012)
8. Drysdale, P.D.: The role of examples in a learner's dictionary. In: Cowie, A.P. (ed.) The Dictionary and the Language Learner. Papers from the EURALEX Seminar at the University of Leeds (1987)
9. Egghe, L., Leydesdorff, L.: The relation between Pearson's correlation coefficient r and Salton's cosine measure. ArXiv e-prints (Nov 2009)
10. Fox, G.: The case for examples. In: Sinclair, J. (ed.) Looking up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary. Collins ELT, London (1987)
11. Fuentes Morán, T., García Palacios, J.: Los ejemplos en el diccionario de especialidad. In: Fuentes Morán, T., García Palacios, J. (eds.) Texto, terminología y traducción. Almar, Salamanca (2002)
12. Humble, P.: The use of authentic, made-up, and controlled examples in foreign language dictionaries. In: Fontenelle, T. (ed.) EURALEX '98. University of Liège, Liège (1998)
13. Kilgarriff, A., Husak, M., McAdam, K., Rundell, M., Rychlý, P.: Gdex: Automatically finding good dictionary examples in a corpus. In: DeCesaris, J. (ed.) 13th EURALEX International Congress. Series del Institut Universitari de Lingüística Aplicada (2008)

14. Kosem, I., Husak, M., McCarthy, D.: Gdex for slovene. In: Proceedings of eLex 2011. pp. 151–159 (2011)
15. Kuguel, I.: La activación del significado especializado. In: Lorente, M., Estopà, R., Freixa, J., Martí, J., Tebé, C. (eds.) Estudis de lingüística aplicada en honor de M. Teresa Cabré Castellví. Series del Institut Universitari de Lingüística Aplicada, Barcelona (2007)
16. Laufer, B.: Corpus-based versus lexicographer examples in comprehension and production of new words. In: Proceedings of the Fifth Euralex International Congress:. pp. 71–76 (1992)
17. Lázaro, J.: El ejemplo en terminología. Caracterización y extracción automática. Ph.D. thesis, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona (2015)
18. Minaeva, L.: Dictionary examples: Friends or foes? In: Tommola, H. (ed.) 5th EURALEX '92 Proceedings. University of Tampere, Tampere (1992)
19. Nesi, H.: The role of illustrative examples in productive dictionary use. Dictionaries: Journal of the Dictionary Society of North America 1(17), 198–206 (1996)
20. Paquot, M.: Exemplification in learner writing: a crosslinguistic perspective. In: Meunier, F., Granger, S. (eds.) Phraseology in Foreign Language Learning and Teaching. John Benjamins Publishing (2008)
21. Sinclair, J.: Collins COBUILD English Language Dictionary. Collins, Birmingham (1987)
22. Spärck-Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 1(28), 11–21 (1972)
23. Volodina, E., Johansson, R., Johansson Kokkinakis, S.: Semi-automatic selection of best corpus examples for swedish: Initial algorithm evaluation. In: Proceedings of the SLTC 2012 workshop on NLP for CALL. pp. 59–70 (2012)
24. Ward Church, K., Hanks, P.: ord association, mutual information and lexicography. Computational Linguistics 1(16), 22–29 (1990)
25. Zgusta, L.: Manual of Lexicography. Walter de Gruyter, Paris (1971)

ISSN 1870-4069

Toward the Design of a Cognitive Tutor for Algebra with Gamification: A Survey of State-of-the-Art

Blanca-Estela Pedroza-Méndez[1], Juan-Manuel González-Calleros[1], Josefina Guerrero-García[1], Carlos-Alberto Reyes-García[2]

[1] Benemérita Universidad Autónoma de Puebla, Puebla, México

[2] Instituto Nacional de Astrofísica Óptica y Electrónica (INAOE), Sta. María Tonantzintla, Puebla, México

{blancaestela.pedroza, jumagoca78, joseguga01}@gmail.com, kargaxxi@inaoep.mx

Abstract. A cognitive tutor is a kind of intelligent tutoring system, which is particularly focused on providing individualized support for the improvement of complex cognitive abilities through the practice of problem resolution. The development of a cognitive tutor involves the analysis of diverse strategies and theories related both to computer science and to pedagogy, as well as the analysis of models related to the discipline in which the tutor will be used. In this paper, an analysis of the state of the art of the topics that have been identified as necessary for the design of a cognitive tutor with gamification to support students with the solving of algebraic problems is conducted. Whenever we talk about an intelligent tutor, models capable of simulating decision making, such as fuzzy models, are required; thus, they are also considered for this analysis. Additionally, comparison tables that allow the identification of opportunity areas for the development of solutions to problems that haven't been thoroughly covered in the literature are presented.

Keywords. Fuzzy cognitive maps, solving algebra problems, cognitive tutors.

1 Introduction

Intelligent Tutorial Systems (ITS) are computer systems designed to facilitate and help a student with the task of learning. They possess experience, for they know the subject to be taught (domain knowledge) and how to teach (pedagogic knowledge), as well as having means to obtain information from the student [1, 2]. A Cognitive Tutor (CT) is a type of ITS with a long-time proven efficacy. Its efficacy is based on its capacity to provide individualized support for the learning of complex cognitive abilities through the practice of problem solving, since they individualize education

through the selection of problems based on a model of the student's current knowledge status, which is updated constantly [3].

One way to complement the design of an intelligent tutorial is by adding gamification features, which consists in the application of principles and elements of game playing in a learning environment with the purposes of influencing behavior, increasing motivation, and promoting student participation [4]. The design and development of a cognitive tutor for helping students develop their skills in algebraic problems solving requires, in its initial steps, the identification of the topics required as a basis, as well as a means to organize all the information required. In order to do this, the 7-module architecture presented by González, Mora & Toledo [1] allows us to classify and construct the state of the art of some of the theories and concepts required for the design and implementation of a cognitive tutor.

This paper is organized as follows: Section 2 provides a brief description of each of the modules of the cognitive tutor and its relation with the main topics and subtopics; In Section 3, some articles related to the main topics are analyzed; and, finally, Section 4 presents, as a result of the analysis, a table in which papers related to intelligent tutorial systems, as well as to one or several of the topics or subtopics covered in Table 1, are classified.

2 General Description of Cognitive Tutor

The Table 1 shows a classification of the modules of a cognitive tutor with their corresponding activities, which, in turn, lead to the identification of topics, concepts and/or theories that, for both practical means and the development of the project, are deemed to be related with the implementation of each module. For example, within the student model module, the student model, the cognitive load theory, and assessment or evaluation, are considered necessary subjects. The student model, as a topic, is important due to the need to identify student preferences throughout the learning process. According to Chrysfiadi et al. [5] it is not effective to assume that all students will follow the same educational model. The student model has two information components: basic information, such as personal data for student identification; and information regarding the knowledge level of the student, as well as his/her cognitive and learning skills. These two information components set the pace for relating the student model to his/her learning style and cognitive skills. Thus, they are considered as subtopics of the student model. Furthermore, the theory of cognitive load, including its two effects - "Expertise Reversal", for the student's level of knowledge, and "Worked Examples, for the handling of the characteristics of the problems that will be presented to the student"-, is considered as a pedagogic foundation. Therefore, the way in which the constant evaluation of the student is conducted is also considered a relevant topic for this module and for the following three: the domain model; the tutor model; and the visualization modules. There are different models for problem solving capability evaluation, for this case, a fuzzy logic model, based on Bloom's Taxonomy, is being used.

Table 1. Description of each module of a cognitive tutor and its related topics.

ITS MODULE	ACTIVITIES [1]	RELATED TOPICS	SUBTOPICS
STUDENT MODEL	• Represents the cognitive state of the student.	Student model	Learning styles
			Cognitive abilities
		Cognitive Load Theory	Expertise Reversal Effect
			Worked Examples Effect
		Assessment or evaluation	Bloom´s Taxonomy
			Fuzzy Models for evaluation
DOMAIN MODEL	• Contains the representation of expert knowledge in areas related to evaluation processes, teaching and learning methodologies.	Fuzzy Logic	Fuzzy Cognitive Maps
			Neuro - Fuzzy networks
		Algebra Teaching.	3UV Model (3 uses of the variable)
		Problem resolution	Heuristics
			Metacognitive strategies
			Belief systems
TUTOR MODEL	• Contains information to decide which tasks are presented to the student, according with the objectives of learning the "domain module"	Instructional Design	Neuro - Fuzzy network
		Feedback	Fuzzy Cognitive Maps
		Assessment or evaluation.	Bloom´s Taxonomy
			Fuzzy Models for evaluation
		Human - Computer Interaction.	Gamification
			Tangible user interfaces
AUTHORING	• Create custom activities	Human - Computer Interaction	Gamification
			Tangible user interfaces
EXECUTION	• Interaction with the student • Knowledge of the game	Feedback	Fuzzy Cognitive Maps
		Human - Computer Interaction	Gamification
			Tangible user interfaces
VISUALIZATION	• Learning analysis. • Game Feedback.	Assessment or evaluation	Bloom´s Taxonomy
			Fuzzy models
		Human - Computer Interaction	Gamification
			Tangible user interfaces
MANAGEMENT	• Management of students • Creation of groups	Human - Computer Interaction	Gamification
			Tangible user interfaces
		Fuzzy Logic	Fuzzy Cognitive Maps

Regarding the domain model module, three topics are considered as fundamental: expert models; algebra teaching; and problem resolution. Expert models are essential and very important for this research because its main objective is related to the implementation of automatized strategies for the proposal of a teaching model through the use of a cognitive tutor. The function of the expert model is to simulate the decision making process that an expert algebra teacher would make, providing the student with problems that correspond to his/her ability for solving them, as well as with suggestions of topics that the student should review. These topics are considered in the tutor model, authorship, and visualization modules. The models used for this decision making process are, precisely, fuzzy cognitive maps which utilizes learning

techniques, in order to train Fuzzy Cognitive Map and choose appropriate weights for its interconnections [6].

Within the tutor model module, the topics to be considered are: instructional design, automatic feedback, evaluation or assessment, and human-computer interaction. Automatic Feedback is a strategy that complements the evaluation and instructional design activities. It is clear that the domain model and tutor model modules should have constant interaction. Both use fuzzy model; however, in the domain model module they are focused on the analysis of the topics and the algebraic problems using fuzzy cognitive maps, whereas the tutor model module is a fuzzy model for student classification based on his/her cognitive ability for problem solving. The authoring, execution and visualization modules consider topics that have been previously described.

3 Analysis of the State of the Art

In order to identify current advances in each of the topics of interest mentioned in Table 1, a review of some related work with these topics is done. The analysis was done considering works mostly from 2012 to date. In each of the subsections of section 3, some works of each of the topics are described.

3.1 Works that are Related with Student Model

In the literature, there are several works related to the automatic detection of student learning styles [7, 8, 9, 10] and the detection of cognitive ability [11], through various computer tools. Nowadays, with the increase of techniques and tools for distance and virtual education, coupled with the growing interest in improving teaching - learning techniques, the generation of models for this purpose continues to emerge. In [11] the authors carry out a study that contributes to the mathematics and education research field by designing calculus tasks to be used as a measure of students' visual and analytic tendencies in calculus. Barron et al [10] also work with detection of learning styles and cognitive abilities through the development of an intelligent tutor system in which identify the emotional state of the student.

3.2 Research Related to Cognitive Load Theory

There are several investigations related to the cognitive load theory, however, within the context of the present investigation, are considered of utmost importance, those that are related, with experiments of the management of the cognitive load, in teaching-learning of mathematics [12, 13], as with the resolution of problems [14, 15, 13] and the management of computational tools [16, 17, 18] that consider as base this theory. By example, in [14] a series of experiments related to the resolution of geometry problems is presented, in which the authors support the importance of having the students work with worked examples. The theory of cognitive load is also used in relation to multimedia, so in [19] conduct a study that identifies how multimedia

interfaces can be developed that do not increase the cognitive load of students, thus preventing the inhibition of learning. Another work in which the authors emphasize the importance of the cognitive load in the teaching process is presented in [12], in which the results of an experiment with visual representations are showed.

3.3 Papers Related to Assessment and / or Evaluation

Although the topic of assessment or evaluation, in most research, are shown as a complementary part of other topics [20, 6, 8, 21, 22] it is important to identify the strategies that are used for evaluation using automated tools. In [23] the authors argue that it is important to develop a form of evaluation that focuses on both the qualitative descriptions of the student's learning process and quantitative information.

3.4 Instructional Design as Objective of some Research

In learning mathematics, a proper instructional design can help the student construct a scheme that will allow him/her to store the information in long-term memory and in turn also help to avoid overloading working memory. An instructional design suitable for the teaching of mathematics could is to enable students to learn real-world problem solving in school mathematics [24, 13, 17, 25]. However, learning/teaching real-world problem solving in school mathematics occurs at different levels along two dimensions: student cognitive functioning and the extent to which teaching is explicit in mediating the real world. Based on this, in [24] propose four-level instructional objectives: 1) Identification, 2) Coordination, 3) Reflection and 4) Transformation.

3.5 Teaching of Algebra

There are several investigations related to the teaching of algebra, some included in projects related to computational tools [3, 26, 27, 28, 29], and others related to pedagogical models [30, 31, 32]. An example is the research carried out in [30], in which a model for classifying algebra problems based on three uses of the variable (as an unknown, as a general number and as a function), is used. With this model, the authors perform a comparative study of problem solving skills with these three uses of the variable, in students from Spain and Mexico. The advances of a project related to the management of video games and techniques of tutorial systems with artificial intelligence, to teach mathematical concepts of algebra are presented in [26].

3.6 Problem Solving as a Topic of Study

The issue of problem solving is very extensive, so there are several studies related to problem solving, which contemplate various strategies, as metacognition [33, 34, 35, 36] and heuristics [37], that have the purpose of analyzing mechanisms to improve the ability of students in this difficult task. Focusing the resolution of problems from metacognition, in the research carried out in [33] the authors conduct research to

identify how metacognition, the beliefs of an individual related to solving mathematical problems and attitudes, affect the process of problem solving.

With regard to investigations that are related to the teaching of algebra and problem solving, reference can be made to the proposals of [38] and [3]. In the first, the authors develop an experiment based on a mathematical theme related to algebraic thinking and model the teaching of algebra using the problem solving approach. In the second the authors develop a cognitive tutor whom they call SimStudent, which has the ability to learn interactively from an expert model through the resolution of directed problems.

3.7 Fuzzy Logic and Educational Systems

In the educational context, models related to cognitive maps are used to generate tools related to the teaching - learning process [39, 6, 40, 5, 41] due to their ability to simulate and predict causal behaviors related to students' learning behavior. In this way, in [39], the authors propose some new dimensions of adaptively like automatic and dynamic detection of learning styles and provides personalization accordingly. It is a literature-based approach in which a personalized adaptive learner model (PALM) was constructed.

3.8 Some Papers on Human Computer Interaction

Gamification and tangible user interfaces are two issues that are generating concerns, within the teaching - learning process, due to the question of whether the strategies used in games and the combination of non - virtual elements with the virtual, help or impair the performance of students in learning. In relation to the use of gamification in education, there are several applications that aim to assess how the gamification elements affect student satisfaction, motivation, enjoyment, learning potency, and degrees over time [42, 43, 26, 1, 27]. Regarding the use of tangible interfaces, in education, there are also several investigations focused on determining if the mixture of real interfaces with virtual systems can help improve the teaching-learning process [44, 45, 46].

4 Summary of Art State Analysis

Finally, as a general synthesis, in Table 2, a list of papers from 2012 to date arranged chronologically are shown, these articles were selected because they include one or more of the topics analyzed in this paper, plus the main theme that has to see with Tutorial Systems. The first column of Table 2 (tagged with R) indicates the reference of the paper, the numbers of the other columns, organized as follows, indicate the topics in Table 1: 1. Tutorial Systems or learning environments. 2. Cognitive Load Theory or Cognitive Theory, 3. Assessment or evaluation, 4. Instructional Design o learning, 5. Algebra or Mathematics Teaching, 6. Problem Solving, 7. Feedback or Suggestions, 8. Human-Computer Interaction, 9. Student Model and 10. Fuzzy sys-

tems or intelligent systems. After the column corresponding to item 10, a column was added indicating the impact of each paper, consulted in September 2017.

Table 2. Classification of papers based on the main topics related with the design of an CT.

R	1	2	3	4	5	6	7	8	9	10	Im	R	1	2	3	4	5	6	7	8	9	10	Im
[39]	x			x						x	1	[15]		x		x	x	x				x	13
[20]			x	x						x	1	[27]	x			x	x	x				x	22
[47]			x	x		x					2	[48]		x								x	48
[44]		x		x	x			x			0	[13]		x		x		x					12
[24]				x		x					0	[40]	x			x					x	x	1
[49]	x			x				x			1	[50]	x			x			x			x	35
[42]				x			x	x			0	[51]	x		x	x	x	x	x				33
[52]	x		x	x						x	1	[53]	x		x	x	x				x		13
[33]				x		x					1	[54]	x			x				x			327
[30]			x	x	x	x					1	[21]			x	x		x				x	1
[38]				x	x	x					3	[36]			x	x	x	x			x		7
[6]			x	x				x		x	4	[28]	x			x	x		x		x		26
[4]			x	x				x			0	[55]		x		x		x					6
[14]		x		x		x					17	[5]	x			x			x			x	20
[19]		x		x				x			5	[56]		x						x	x		18
[45]				x				x			9	[57]			x	x				x			195
[37]		x		x							3	[16]	x	x	x	x					x	x	9
[43]			x	x				x			24 2	[29]	x			x	x			x		x	41
[58]				x	x	x					4	[17]	x	x		x	x						10
[3]	x			x	x	x				x	36	[22]	x		x	x		x					27
[59]	x			x				x			0	[25]		x		x							20
[60]			x	x		x	x				31	[61]		x		x							15
[62]				x						x	1	[63]	x	x		x	x						25
[12]		x		x	x			x			10	[46]	x			x	x			x			28
[31]				x	x	x					3	[9]	x			x					x	x	1
[34]	x			x	x			x			46	[18]		x		x				x			9
[64]	x			x							0	[10]	x			x	x					x	7
[65]		x		x							63	[66]		x	x	x							0
[67]		x						x		x	0	[68]		x		x				x			22
[26]				x	x			x			7	[69]	x			x					x		16
[7]	x			x					x	x	22	[2]	x			x					x	x	19
[1]	x			x				x			15	[70]		x		x							51
[8]			x	x					x	x	26	[71]	x			x					x	x	88
[32]				x	x						6	[72]		x		x	x	x					37
[73]									x	x	104	[74]			x	x						x	67
[75]	x			x	x					x	12	[76]		x		x				x			161
[35]				x	x	x		x			1	[77]	x	x		x			x	x			14

4.1 Results of Analysis of Classification of Papers

One of the main themes of this review of the state of the art is the ITS, but if we include the part of the pedagogical sustenance, the subject of CLT should be considered, analyzing the classification of Table 2, it is observed that 40% of the articles have are related to ITS and 33% with the subject of CLT, but what really interesting to visualize the areas of opportunity in the development of ITS with cognitive pedagogical sustenance are the works that include these two themes, of which they are only 5%, which means that there is a large area of opportunity to develop projects that include these two themes. One of the topics included in 95% of the articles is instructional design or teaching learning, evidently because the main applications of both ITS and CLT are related to teaching - learning. Regarding the topics that each article contains, the classification is as follows: 18% is related to 2 of the 10 topics, 35% with 3, 33% with 4, 11% with 5 and only 3% includes 6 of the 10 topics, that is, no

paper considers more than 6 topics, so that developing research that includes 7,8,9 or all themes are also areas of opportunity for future research projects.

5 Conclusions

An analysis of the conceptual topics related to the design and development of a cognitive tutor focused on the teaching of algebra has been conducted. The topics have been classified according to the module of the intelligent tutorial system to which they belong. A total of ten topics have been chosen, and a state of the art analysis has been conducted for each one of them.

We have been able to identify that some topics have been thoroughly studied, and that these studies are truly useful tools.

Regarding the modules that constitute a cognitive tutor, it has been observed that each one of them has its particular degree of difficulty due to the topics that have to be mastered. Nevertheless, according to the classification shown in Table 1, when the student model, domain model and tutor model modules are implemented, the cognitive tutor can be considered as implemented since the remaining modules are related to topics that have already been considered in the aforementioned modules.

Finally, the construction of a cognitive tutor requires from the designer and the developer an interdisciplinary formation, both in the knowledge field that will be the training object, and in artificial intelligence and computer science methods and techniques. All these factors pose a great challenge.

References

1. González, C., Mora, A., Toledo, P.: Gamification in intelligent tutoring systems. In: Proc. Second Int. Conf. Technol. Ecosyst. Enhancing Multicult. - TEEM ’14, 221–225 (2014)
2. Fazel, M. H., Khademian, M., Minaei, B., Türkşen, I.B.: A Fuzzy Expert System Architecture for Intelligent Tutoring Systems: A Cognitive Mapping Approach. J. Intell. Learn. Syst. Appl., vol. 4, no. February, 29–40 (2012)
3. Matsuda, N., Cohen, W.W., Koedinger, K.R.: Teaching the teacher: Tutoring simstudent leads to more effective cognitive tutor authoring. Int. J. Artif. Intell. Educ., 25(1), 1–34 (2015)
4. Bossomaier, T.: Serious Games and Gaming, Vol 4, 201–232 (2015)
5. Chrysafiadi, K., Virvou, M.: A knowledge representation approach using fuzzy cognitive maps for better navigation support in an adaptive learning system. Springerplus, 2(1), 81–94 (2013)
6. Bolivar, H. B., González, R. C., Pascual, J. E., Sanjuán, O. M.: Assessment of learning in environments interactive through fuzzy cognitive maps. Soft Comput, 19, 1037–1050 (2015)
7. Feldman, J., Monteserin, A, Amandi, A.: Automatic detection of learning styles: state of the art. Artif. Intell. Rev., no. May 2014, 157–186 (2014)
8. Jegatha, L. D., Baskaran, R., Kannan, A.: Learning styles assessment and theoretical origin in an E-learning scenario: a survey. Artif. Intell. Rev., 42(4), 801–819 (2014)
9. Zatarain, R. C., M., Barrón, L. E., Olivares, L. C., Reyes, C. G.: Integrating learning styles

and affect with an intelligent tutoring system. Proc. - 2013 12th Mex. Int. Conf. Artif. Intell, MICAI 2013, 247–253 (2013)
10. Barron., L. E., Zatarain, R. C., Beltrán, J. V., Cibrian, F. R., Hernández, Y. P.: An Intelligent and Affective Tutoring System within a Social Network for Learning Mathematics. 651–661 (2012)
11. Haciomeroglu, E. S.: The role of cognitive ability and preferred mode of processing in students' calculus performance. Eurasia J. Math. Sci. Technol. Educ., 11(5), 1165–1179 (2015)
12. Yung, H., Paas, F.: Effects of computer-based visual representation on mathematics learning and cognitive load. Educ. Technol. Soc., 18(4), 70–77 (2015)
13. Ngu, B. H, A. Yeung, A. S., Tobias, S.: Cognitive load in percentage change problems: Unitary, pictorial, and equation approaches to instruction. Instr. Sci., 42(5), 685–713 (2014)
14. Chen, O., Kalyuga, S., Sweller, J.: The worked example effec, the generation effect and element interactivity. J. Educ. Psychol., 107(3), 689–704 (2015)
15. Lin J. J., Lin, S. S.: Cognitive Load for Configuration Comprehension in Computer-Supported Geometry Problem Solving: An Eye Movement Perspective. Int. J. Sci. Math. Educ., 12 (3), 605–627 (2014)
16. Lach, P.: Student's Effort During Assessment. 346–351 (2013)
17. Reed, S. K., Corbett, A., Hoffman, B., Wagner, A., MacLaren, B.: Effect of worked examples and Cognitive Tutor training on constructing equations. Instr. Sci., 41(1), 1–24 (2013)
18. Andrade-Lotero, L. A.: Teoría de la carga cognitiva, diseño multimedia y aprendizaje: Un estado del arte. Magis, 5(10), 75–92 (2012)
19. Cheng, T. S., Lu, Y. C., Yang, C. S.: Using the multi-display teaching system to lower cognitive load. Educ. Technol. Soc., 18(4), 128–140 (2015)
20. Goksu, I.: The Evaluation of the Cognitive Learning Process of the Renewed Bloom Taxonomy Using a Web Based Expert System. 15(4), 135–152 (2016)
21. Sharma, M., Chawla, S.: A Tools for Creating Constructivist Learning Environment and Assessing Knowledge Development using Concept Maps. 5(8) (2014)
22. Reimann, P., Kickmeier-Rust, M., Albert, D.: Problem solving learning environments and assessment: A knowledge space theory approach. Comput. Educ., 64, 183–193 (2013)
23. Hassan, O. B.: Learning theories and assessment methodologies - an engineering educational perspective. Eur. J. Eng. Educ., vol. 36, no. March 2015, 327–339 (2011)
24. Jurdak, M.: Learning and Teaching Real World Problem Solving in School Mathematics. 181–195 (2016)
25. Rey, G. D., Andreas, F.: The expertise reversal effect concerning instructional explanations. Instr. Sci., 41(4), 635–656 (2013)
26. Faghihi, U., Brautigam, A., Jorgenson, K., Martin, D., Brown, A., Measures, E.: Maldonado-Bouchard, S.: How gamification applies for educational purpose specially with college algebra. Procedia Comput. Sci., 41, 182–187 (2014)
27. Long, Y., Aleven, V.: Gamification of Joint Student/System Control Over Problem Selection in a Linear Equation Tutor. Science (80-87), 37, 849–874 (2014)
28. Walker, E., Rummel, N., Koedinger, K. R.: Adaptive intelligent support to improve peer tutoring in algebra. Int. J. Artif. Intell. Educ., 24(1), 33–61 (2014)
29. Mavrikis, M., Noss, R., Hoyles, C., Geraniou, E.: Sowing the seeds of algebraic generalization: Designing epistemic affordances for an intelligent microworld. J. Comput. Assist. Learn., 29(1), 68–84 (2013)
30. Álvarez, I., Gomez-Chacón, I. Ursini, S.: Understanding the Algebraic Variable: comparative Study of Mexican and Spanish Students. 11(6), 1507–1529 (2015)

31. Zeljić, M.: Modelling the Relationships Between Quantities: Meaning in Literal Expressions. 11(2), 431–442 (2015)
32. Jupri, A., Drijvers, P., Heuvel-Panhuizen, M.: Student difficulties in solving equations from an operational and a structural perspective. Int. Electron. J. Math. Educ., 9(1-2), 39–55 (2014)
33. Bas, F., Sagirli, M.: The Metacognitive Awarenesses of Pre- Service Secondary School Mathematics Teachers, Beliefs, Attitudes on Problem Solving, and Relationship Between Them. 12(2), 464–482 (2016)
34. Arroyo, I., Woolf, B. P. Burelson, W., Muldner, K., Rai, D., Tai, M.: A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. Int. J. Artif. Intell. Educ., 24(4), 387–426 (2014)
35. Kramarski, B., Friedman, S.: Solicited versus Unsolicited Metacognitive Prompts for Fostering Mathematical Problem Solving Using Multimedia. J. Educ. Comput. Res., 50(3), 285–314 (2014)
36. Tzohar-Rozen, M., Kramarski, B.: Metacognition, Motivation and Emotions: Contribution of Self-Regulated Learning to Solving Mathematical Problems. Glob. Educ. Rev., 1(4), 76–95 (2014)
37. Eisenmann, P., Novotná, J., Přibyl, J., Břehovský, J.: The development of a culture of problem solving with secondary students through heuristic strategies. Math. Educ. Res. J., 27, (4), 535–562 (2015)
38. Bailey, J., Taylor, M.: Experiencing a mathematical problem-solving teaching approach: Opportunities to identify ambitious teaching practices. 17, 111–124 (2015)
39. Sweta, S., Lal, K.: Personalized Adaptive Learner Model in E-Learning System Using FCM and Fuzzy Inference System. Int. J. Fuzzy Syst. (2017)
40. Peña-Ayala, A., Sossa-Azuela, J. H.: Decision Making by Rule-Based Fuzzy Cognitive Maps: An Approach to Implement Student-Centered Education. Fuzzy Cogn. Maps Appl. Sci., 54, 107–120 (2014)
41. Zouhair, A., En-Naimi, E. M., Amami, B., Boukachour, H., Person, P., Bertelle, C.: Intelligent tutoring systems founded on the multi-agent incremental dynamic case based reasoning. Cist 2012 - Proc. 2012 Colloq. Inf. Sci. Technol., 74–79 (2012)
42. Mozelius, P., Fagerström, A., Söderquist, M.: Motivating Factors and Intrinsic Integration of Knowledge in Educational Games. Eur. Conf. Games Based Learn., 500–508 (2016)
43. Hanus, M. D., Fox, J.: Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. Comput. Educ., 80. 152–161 (2015)
44. Guerrero, G., Ayala, A., Mateu, J., Casades, L., Alamán, X.: Integrating Virtual Worlds with Tangible User Interfaces for Teaching Mathematics: A Pilot Study. Sensors, 16(11), 1775 (2016)
45. Cuendet, S., Dehler-Zufferey, J., Ortoleva, G., Dillenbourg, P.: An integrated way of using a tangible user interface in a classroom. Int. J. Comput. Collab. Learn., 10(2) 183–208 (2015)
46. Starcic, A., I., Cotic, M., Zajc, M.: Design-based research on the use of a tangible user interface for geometry teaching in an inclusive classroom geometry teaching in an inclusive classroom. 44(5), 729–745 (2013)
47. Grover, S., Bienkowski, M., Niekrasz, J., Hauswirth, M.: Assessing Problem-Solving Process a t Scale. 245–248 (2016)
48. Gray, S. A., Zanre, E., Gray, S. R..: Fuzzy Cognitive Maps for Applied Sciences and Engineering. Intell. Syst. Ref. Libr. 54, 159–175 (2014)
49. Millis, K., Forsyth, C., Wallace, P., Graesser, A., Timmins, G.: The Impact of Game-Like

Features on Learning from an Intelligent Tutoring System. Technol. Knowl. Learn., 22(1) 1–22 (2016)

50. Rivers, K., Koedinger, K., R.: Automating Hint Generation with Solution Space Path Construction. 329–339 (2014)
51. Roll, I., Baker, R., Aleven, V., Koedinger, K., R.: On the Benefits of Seeking (and Avoiding) Help in Online Problem-solving Environments. J. Learn. Sci., 23(4), 537–560 (2014)
52. Ramirez-Noriega, A., Reyes, J. R., Martinez-Ramirez, Y., Jimenez, S., Inzunza, S.: Using Bayesian Networks for Knowledge Representation and Evaluation in Intelligent Tutoring Systems. Data Min. Acad. Databases, (2), 189–198 (2016)
53. San Pedro, M. O., Baker, R., Rodrigo, M. M.: Carelessness and affect in an intelligent tutoring system for mathematics. Int. J. Artif. Intell. Educ. 24(2), 189–210 (2014)
54. Seaborn, K., Fels, D. I.: Gamification in theory and action: A survey. Int. J. Hum. Comput. Stud., 74, 14–31 (2014)
55. Alzahrani, I., Woollard, J.: The Role of the Constructivist Learning Theory and Collaborative Learning Environment on Wiki Classroom, and the Relationship between Them. Online Submiss (2013)
56. Francisco-Aparicio, a., Gutiérrez-Vela, F. L., Isla-Montes, J. L., González-Sánchez, J. L: Gamification: Analysis and Application. 45–61 (2013)
57. Kulkarni, C., Papadopoulos, K., Cheng, J., Koller, D., Klemmer, S. R.: Peer and Self-Assessment in Massive Online Classes. 20(6) (2013)
58. Hong, J. Y., Kim, M. K.: Mathematical Abstraction in the Solving of Ill-Structured Problems by Elementary School Students in Korea. EURASIA J. Math. Sci. Technol. Educ., 12(2), 267–281 (2015)
59. Mukhtar, H.: Conceptual Model to Incorporate Serious Games Mechanics in Intelligent Tutoring Systems. 5, 13–18 (2016)
60. Piech, C., Huang, J., Sahami, M., Guibas, L.: Autonomously generating hints by inferring problem solving policies. 195–204 (2015)
61. Roelle, J., Berthold, K.: The expertise reversal effect in prompting focused processing of instructional explanations. Instr. Sci., 41(4), 635–656 (2013)
62. Bodyanskiy, Y. V., Tyshchenko, O. K., Deineko, A. O.: An Evolving Neuro-Fuzzy System with Online Learning/Self-learning. Int. J. Mod. Educ. Comput. Sci., 7(2), 1–7 (2015)
63. Schwonke, R., Ertelt, A., Otieno, C., Renkl, A., Aleven, V., Salden, R. J.: Metacognitive support promotes an effective use of instructional resources in intelligent tutoring. Learn. Instr., 23(1), 136–150 (2013)
64. Caro, M. F., Jimenez, J. A.: MOF-based metamodel for pedagogical strategy modeling in Intelligent Tutoring Systems. In: 2014 9th Computation Colomb. Conferences 9CCC 2014, 1–6 (2014)
65. Choi, H. H., VanMerriënboer, J. J., Paas, F.: Effects of the Physical Environment on Cognitive Load and Learning: Towards a New Model of Cognitive Load. Educ. Psychol. Rev., 26(2), 225–244 (2014)
66. Chen, Y.: Assessment System based on Scaffolding and Bloom' s Theory. J. Digit. Content Technol. its Appl., 6(6), 218–227 (2012)
67. Conde-ram, J. C., Abraham, S.: Evocación de hábitos en personajes virtuales mediante Mapas Cognitivos Difusos y técnicas de videojuegos. 73, 73–87 (2014)
68. Cheon, J., Grant, M. M.: The effects of metaphorical interface on germane cognitive load in Web-based instruction. Educ. Technol. Res. Dev., 60(3), 399–420 (2012)

69. Dung, P. Q., Florea, A. M.: A literature-based method to automatically detect learning styles in learning management systems. In: Proc. 2nd Int. Conf. Web Intell. Min. Semant. - WIMS '12, 1 (2012)
70. Kalyuga, S., Rikers, R., Paas, F.: Educational Implications of Expertise Reversal Effects in Learning and Performance of Complex Cognitive and Sensorimotor Skills. Educ. Psychol. Rev., 24(2), 313–337 (2012)
71. Latham, A., Crockett, K., McLean, D., Edmonds, B.: A conversational intelligent tutoring system to automatically predict learning styles. Comput. Educ., 9(1), 95–109 (2012)
72. Leppink, J., Broers, N. J., Imbos, T., vanderVleuten, C. P., Berger, M. P.: Self-explanation in the domain of statistics: An expertise reversal effect. High. Educ., 63(6), 771–785 (2012)
73. Kar, S., Das, S., Ghosh, P. K.: Applications of neuro fuzzy systems: A brief review and future outline. Appl. Soft Comput., 15, 243–259 (2014)
74. Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., Levy, R.: Design and discovery in educational assessment: Evidence-centered design, psychometrics, and Educational Data Mining. J. Educ. Data Min., 4(1), 11–48 (2012)
75. Khachatryan, G. A., Romashov, A. V., Khachatryan, A. R., Gaudino, S. J., Khachatryan, J. M., Guarian, K. R., Yufa, N. V.: Reasoning mind genie 2: An intelligent tutoring system as a vehicle for international transfer of instructional methods in mathematics. Int. J. Artif. Intell. Educ., 24(3), 333–382 (2014)
76. Paas, F., Sweller, J.: An Evolutionary Upgrade of Cognitive Load Theory: Using the Human Motor System and Collaboration to Support the Learning of Complex Cognitive Tasks. Educ. Psychol. Rev., 24(1), 27-45 (2012)
77. Paquette, L., Lebeau, J. F., Andre, M.: Automating Next-Step Hints Generation Using ASTUS. Springer- Verlag Berlin Heidelberg, 201–211 (2012)

ISSN 1870-4069

Gathering, Classifying and Visualizing Results from Social Surveys using OCR and Machine Learning Techniques

David Céspedes-Hernández, Juan Manuel González-Calleros, Josefina Guerrero-García, Liliana Rodríguez-Vizzuett

Benemérita Universidad Autónoma de Puebla, Puebla, Mexico

{dcespedesh, juan.gonzalez}@cs.buap.mx, {jguerrero, lilianarv}@cs.buap.mx

Abstract. Statistical representation of data as graphics is useful for making its interpretation easier, and for helping on decision making processes. When it comes to information that is obtained as result of social surveys, the problem is not only on presenting it, but on gathering the results by different means, and classifying them into categories, considering that for open questions, this is a process that is commonly done in a manual way. This paper is aimed on presenting a software tool for gathering data from social surveys relying on OCR technology or through a web application, classifying it using Machine Learning techniques, and presenting the resulting information in a mostly automatic way. For illustrative purposes, a case of study of a survey that was applied in Puebla, Mexico, in 2013 is considered. Quantitative results were obtained and reported for the classification process, and general user interfaces of the developed application are presented and described.

Keywords. Machine learning, information visualization, social survey, web development, optical character recognition.

1 Introduction

When gathering data for research purposes, surveys are one of the most commonly used methods [15]. Several authors have addressed the creation and application of surveys for obtaining data about specific subjects from social, psychological and statistical points of view to mention a few. Also, it is remarkable that to a number of domains such as politics, economics and marketing, surveying is relevant [2, 10, 17].

Usually, it is not possible to survey all of the members within a population and in order to have representative results from surveys, it is necessary to carefully pick samples that reflect tendencies of how a determined population thinks or acts. In this sense, if it is of interest to understand the opinion of a complete town towards some concept, it will be necessary to determine a sample size [3], to survey the members of the sample (traditionally but not limited to paper based questionnaires), to gather the

results, and to obtain and present information from them through a diversity of statistical methods.

However, for some studies such as census, or surveys on social needs applied on remote communities, because of the sample size and its geographic distribution, gathering the results, organizing them, and giving an interpretation is not always a straightforward process.

In addition, to process surveys that were applied to large samples can be a time consuming, expensive task, and human errors may happen either when gathering the data, when capturing it (if needed), when categorizing the resulting information, or when generating statistics for analysis.

It is evident that multiple domains are affected by this problem, and that its solution through a software tool would help in minimizing and quantifying the errors, reducing the invested time and the cost of the whole process, and presenting statistical information in an opportune way.

The general objective of this work, is to develop an application for gathering results either from paper based surveys or from digitally stored questionnaires, classifying such results into domain specific categories, and finally giving visual representation to that information for its further analysis. As it can be inferred, this objective can be decomposed into a set of specific goals to be accomplished:

1) Support data gathering from paper based surveys.
2) Support data gathering from digitally stored questionnaires.
3) Classify the gathered data into a set of defined, specific to the domain categories.
4) Present the classified information in a graphic way for its analysis.

The remainder of this paper is structured into sections. Section two contains the state of the art regarding techniques and tools that are useful for the development of this work: Optical Character Recognition tools, Machine Learning techniques applied to text classification, and suites or Application Programming Interfaces (APIs) for information visualization. The third section is aimed on describing the development of an application for meeting the objectives that were defined in previous lines. Later, results of the implementation of the developed application are reported in Section 4.

Next, Section 5 consists on a discussion about the before presented results, comparing them to the usage of different algorithms for text classification on the proposal context. Finally, in Section 6, conclusions and future work are provided in terms of how the defined goals were accomplished and of further research to be done.

2 State of the Art

Considering the specific objectives that were described in the Introduction, it is possible to identify the requirement of exploring techniques and tools that allow to address them. This section is dedicated to describe a state of the art on such concepts. First of all, taking into account that it is relevant for this work to have a mechanism for automatically processing paper based questionnaires, Optical Character Recognition (OCR) is presented. Next, as after having a set of answers either from paper based or

digital surveys it will be needed to classify such information, some Machine Learning (ML) algorithms for text classification are described. And Finally, in order to meet the requirement of presenting the resulting data in a graphic way, suites and tools for Information Visualization (IV) are studied.

2.1 Optical Character Recognition

OCR is not a new problem in the field of pattern recognition. Moreover, this has been addressed by several authors and from different perspectives for decades. The definition of the term and the first works on the matter date back to 1930's [9, 20] and specifically involving computers, to the decade of 1950's [16].

The earlier attempts of OCR algorithms reported a precision of around 60% correct recognitions, while modern proposals are accurate in a 97 to 100% [4], and what is pursued nowadays, is to improve the processing time, the availability of recognition services, and other add-ons such as variety of formats of the outcome, APIs definition, and real-time translation services.

As providing a complete overview on OCR tools and algorithms is not an objective of this paper, a summarized comparison is reported in Table 1, taking only into account minimum characteristics such as proposal name, whether if an SDK/API is provided for it or not, the format of its outcome, and if multilingual support is given.

Table 1. Summarized comparison of OCR tools.

\	**SDK/API provided?**	**Outcome format**	**Multi lingual support**
Google Vision[1]	Yes	JSON structure.	Yes
Omnipage[2]	Yes	XML and text files.	Yes
Online OCR[3]	Yes	Text and Office suite formats.	Yes

For this particular work, the proposal by Google Cloud Platform is especially interesting as it provides an API that allows its inclusion to web applications, it provides high precision detection (even for handwritten texts), it allows to perform up to 1,000 free requests monthly, and it counts on active support and documentation by Google, and by its developers' community.

2.2 Machine Learning Techniques Applied to Text Classification

By definition, Text Classification (TC) is the assignment of Boolean values to each pair (di, ci) ϵ D X C, where D represents the documents to be classified, and C is the set of classes in which the documents will be categorized [19].

[1] https://cloud.google.com/vision/
[2] https://www.nuance.com/
[3] https://www.onlineocr.net/

Since the early 60's, TC has become an activity of interest for its application in various contexts that require information retrieval such as news filtering and organization, opinion mining, and email classification (spam filtering), among others. On the first efforts about TC, a set of rules was manually defined for encoding expert knowledge on how document classification was supposed to be done. Then, those rules were used in order to classify incoming data [11]. The main disadvantage then, was the dependence towards experts for the creation of constrains, and that the accuracy was directly proportional to the extension and completeness of the set of rules.

During the decade of 1990, the Knowledge Engineering community worked on the application of ML techniques for providing a more automatic and more accurate approach, following the principle of making this process less reliant on the expert fed.

A number of ML techniques have been designed or adapted for accomplishing TC. If well most methods are created for being used on the quantitative data related to text, such approaches are usually applied depending on the characteristics of the documents to be classified, delivering different results in terms of precision, computational cost, and usage or implementation complexity.

The most commonly used proposals for TC in the literature, are as follows: Decision trees, Pattern-based classifiers, SVM classifiers, Neural network classifiers, and Bayesian classifiers [12].

In general, probabilistic classifiers, also called generative classifiers, are designed for using a mixture model for generation of the underlying documents. This mixture model typically assumes that each class is a component of the mixture. Each mixture component is essentially a generative model, which provides the probability of sampling a particular term for that component or class [1].

According to the nature of the gathered data, to be described in Section 3.1, Probabilistic Bayesian methods are considered for the elaboration of this proposal. Bayesian classifiers, are probabilistic techniques, based on modeling the underlying word features in different classes. Their operation is based on classifying text considering the posterior probability of the documents belonging to different classes, taking into account the presence of the word that was modeled in the documents [13].

2.3 Suites and APIs for Information Visualization

IV is the process of transforming abstract information into more easily understandable, graphical forms [18]. Typically, before applying statistical graphics or other representation techniques, it is necessary to preprocess data i.e. converting non-numerical data into a numerical form, and organizing or arranging information, for easier understanding and analysis.

Nowadays, along with the development of systems that allow to work with large amounts of information, and to deliver results or reports for analysis or for supporting decision making processes, the need of tools for visualizing such information is imminent [6]. With that necessity, came the proposal by the research community on the matter, of providing a diversity of tools, techniques and applications.

For this work, it is of especial interest those proposals having an application on web environments, providing the options of using different representations for numer-

ical data, counting on support by developers' community, and allowing the creation of custom graphs. In this sense, Google Charts API[4], Chart.js [5], JFreeChart [8], and InfoViz[5] were considered and contrasted. It was finally decided to use InfoViz library because of the variety of charts that it provides, and that they may be useful for the purposes of this proposal, and for the future work to be described in Section 6.

3 Development of a Software Tool for Gathering, Classifying, and Visualizing Results from Social Surveys

Taking into account the objectives that were defined in the introduction and applying what was reported on the state of the art towards their accomplishment, it was decided to create a software tool for gathering results from social surveys, that allows to classify such results in a semiautomatic way, and that reports information about the done classification in a graphic manner.

It was also noticed, that due to the nature of the surveys, the gathering process should consider that data may come either from scanned paper based questionnaires, or from digital versions of such questionnaires.

In order to expose the development that was done, a case of study is considered. This experiment consisted on taking data from 25,014 answered surveys that were applied in the city of Puebla, Mexico during the year 2013. The objective of that research work was to understand the population opinion regarding the government performance, and to identify needs in terms of *security*, *health*, *employment*, *economy*, and *education*.

The inquiry form consisted on 40 questions divided on 3 blocks. The first set of questions was defined for getting relevant information about the context of the respondent. From this point, it was possible to get sociodemographic data.

The second block contained only closed questions, and was aimed on getting quantitative data representing the evaluation from the population to the government. Finally, the third block was designed for obtaining qualitative information regarding what the population identifies as their specific needs.

The methodology for this study consisted on surveying people from different locations within the city, as an attempt to have a representative sample, using a paper based version of the questionnaire as the one depicted in Figure 1.

Every day, the answered surveys were sent to a capture department in order to get the results digitalized. At the same time, data capturers had the assignment of classifying the answers from block three into the five previously mentioned categories.

Once the determined period of time for applying the surveys was ended, and all the answers were computed, a tailored web application was implemented for counting, organizing, and presenting the resulting information in a graphic way.

From this experience, some insights were taken:

[4] https://google-developers.appspot.com/chart/
[5] http://infoviz.org/

a) It was necessary to involve a big team for surveying people, and another big team for capturing and categorizing the data, making this an expensive experiment.
b) Processing a survey, took around 25 man-minutes each (for the 25,014 questionnaires it took 10,400 man-hours).
c) During the data capture and manual classification process, human errors happened.
d) Even though the capture department completed the classification process and therefor acquired experience, if it was necessary to add more surveys to the study i.e. include more results, or to start a new study with same parameters, the time for accomplishing the tasks would not significantly decrease.

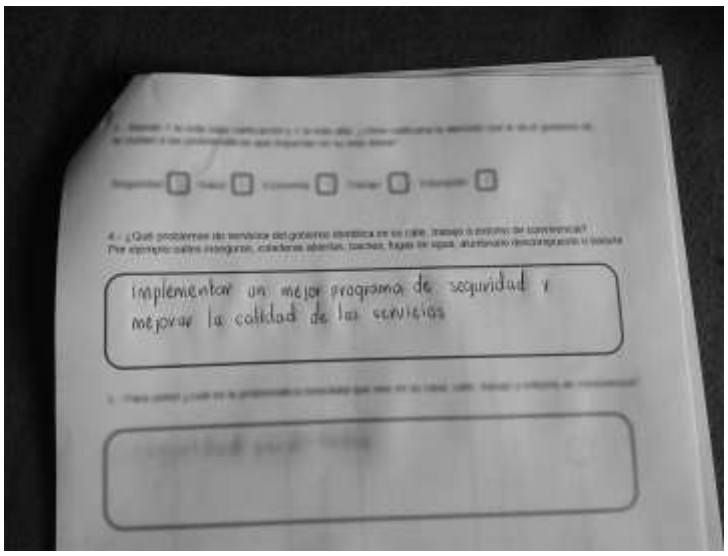

Fig. 1. Extract of a paper based questionnaire used on the case of study.

Next sections are dedicated to describe the development of a software web-based application for reaching the objectives that were defined for this work, instantiating it in terms of the introduced case of study.

3.1 Data Gathering

In the presented case of study, paper based surveys were used. This caused the necessity of capturing the answers in a separate process.

Aligned with the objectives of this work, our proposal for gathering data in a more efficient way, consists on either developing a simple web based application, and to use it for directly entering the information in an electronic format, or to scan the questionnaires and use OCR tools for its digitalization.

Regarding the provision of a system for enabling the application of digital surveys, Figure 2 shows an extract of its User Interface (UI). The use of this tailored application is pretty straightforward, as each question is listed along with a component for

inputting the response, and after completing each of the questionnaire blocks, it is possible to save changes into a local database.

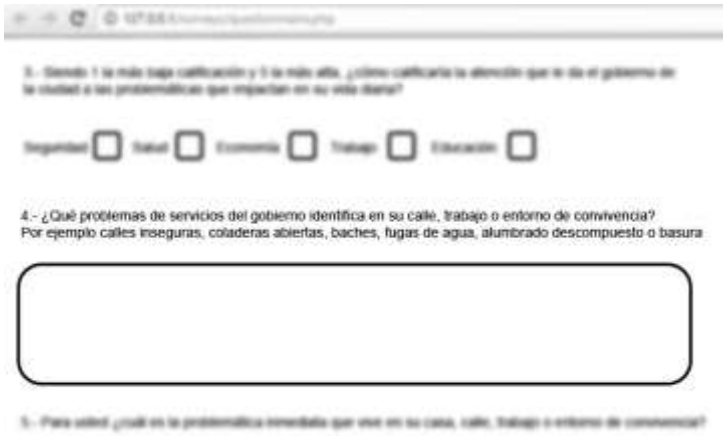

Fig. 2. User interface of the developed survey system.

An interesting feature of this software tool, is that it allows to apply the surveys off-line, and later, when Internet connection is available, to perform the information load to a data base stored in a proprietary server.

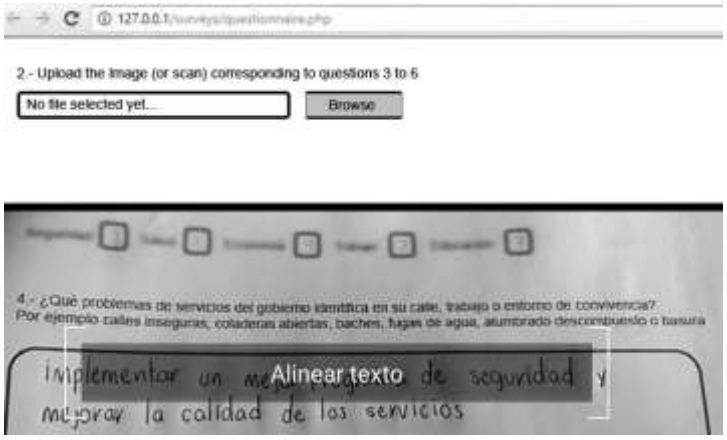

Fig. 3. User interface of the developed OCR application.

The second alternative given for data gathering, consists on providing a software tool for allowing document scanning, so later, by using OCR, digitalize the responses. For this purpose, it was necessary to develop another tool, which UIs are shown in Figure 3.

The interaction with this software, consists on following the instructions in the screen, for uploading/scanning the paper based surveys in a specific order. The out-

come of Google Vision OCR, presented in Figure 4, is a JSON format structure with the read text and vectors containing the position of each word on the image.


```
"responses": [
  {
    "textAnnotations": [
      {
        "locale": "es",
        "description": "implementar un mejor programa
        de seguridad\ny mejorar la calidad de los
        servicios\n",
        "boundingPoly": {
          "vertices": [
            {
              "x": 26,
              "y": 47
            },
            {
              "x": 217,
              "y": 47
            },
            {
```


Fig. 4. JSON outcome from Google's OCR.

As in the other option for gathering information, part of the process (image capture or scanning) can be done off-line, and then, the file upload should be performed when having Internet connection.

The above described systems address the survey and data capturing processes. The next step in the methodology is then the classification of such data into the already mentioned five categories of interest. In the following section, a ML approach for that purpose is presented.

3.2 Data Classification

The manual classification that is described in the case of study, followed two main principles: 1) All responses must fit in a category, and 2) Every answer can only be labeled with one class, even if two or more are applicable. Following those same principles, and considering that manual classification has already been done, it was possible to instantiate a probabilistic Bayesian classifier.

Before implementing the classifier, it was necessary to preprocess information and to prepare its parameters. First of all, the structure of the information was analyzed, and it was noticed that most responses were not longer than 10 words, and that there were even some of them containing only one. Taking this into account, documents were normalized, i.e. accents and punctuation were removed, and all text was converted to lowercase.

After normalization, documents were separated according to the category they belong to, in order to calculate the frequency of meaningful words for each class. As it can be inferred, in this counting closed words were not considered.

The outcome of this operation consisted on five lists of words, one for each category, along with their appearance in the set of documents. The 50 most frequent words

were selected from each one and then combined, removing repeated utterances, and resulting on a 149 words dictionary organized in an array of 149 Strings (categories).

Using that dictionary, each document was translated to numerical comma separated data, considering if whether a word from the dictionary appeared in the document or not, and how many times it was found in there, respecting the already given classification (class), as it can be seen in Figure 5.

```
1  implementar un mejor programa de seguridad
2  y mejorar la calidad de los servicios
3
4  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,
5  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
6  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
7  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
8  1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
9  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,
10 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,7
```

Fig. 5. Example of document translation to numerical comma separated form, using the created dictionary.

With such defined categories and classes, the Bayesian classifier implementation was possible. The execution of this algorithm is transparent to the user, but for research purposes metrics can be applied for evaluating its precision and performance. The results obtained from the training and testing of this approach are presented in Section 4.

Once training of the classifier is complete, and the organization of the testing set is done, it is possible to automatically classify new coming information and to move on to the next stage of the methodology, enable visualization of the obtained information. The development of the IV module is briefly described in the next section.

3.3 Information Visualization

According to the methodology followed in the case of study, and the objectives that were established in the introduction, IV is the last phase of the process. The idea that is pursued with linking the information presentation to the data classification, is to allow users to have real-time interpretation of the gathered data.

The development of this module, consisted on creating a generic application, that given a set of classes, a chart type, and a data source, prepares a UI for displaying graphics corresponding to the fed information using InfoViz.

One of the resulting UI for the case of study is presented in Figure 6. For this specific example, the used classes are: *security*, *health*, *economy*, *education*, and *employment*; the chart type is pie chart; and the data source is the outcome of the classification process.

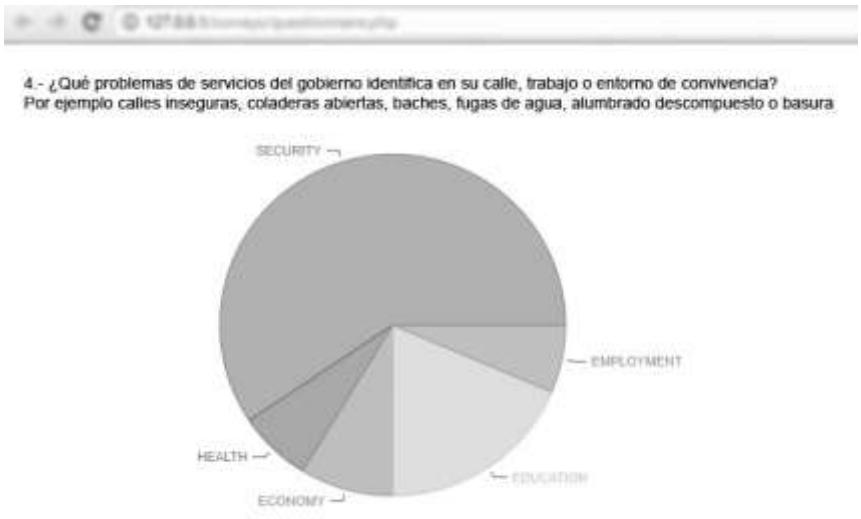

Fig. 6. User interface of the information visualization module.

This graphic presentation, combined with the ones of other questions, supports the interpretation of the study, and a decision making process as it was the objective of the research experiment. The following section is dedicated to detail the results that were obtained during the development of this work.

4 Results

The present work had results in different terms. During its development, a web based application was created for allowing users to survey, gather data from surveys, classify information, and present it in a graphic way.

For the implementation of such software, along with web development languages and tools such as HTML, PHP, JavaScript, jQuery, MySQL, Google's OCR API, a Bayesian classifier, and InfoViz Library were used. Regarding the Bayesian classifier that was used, it was possible to gather statistic information about its performance. Table 2 contains the confusion matrix that was from it obtained, after using 65% of data for training and 35% as testing set.

From the 25,014 documents that were available, 21,293 were successfully classified, representing a total error of 14.88%. If well this number provides information regarding the behavior of the classifier, it is necessary to carry out a deeper analysis to understand its limitations.

There were 13,236 documents classified into the security class. In the implementation of the classifier, 95.58% (12,652/13,236) of such items were correctly organized. In a similar way, the precision for the other classes were 72.97% (1,461/2,002), 58.40% (1,856/3,178), 81.40% (3,961/4,866), and 78.69% (1,363/1,732).

The class that got the lowest precision is Economy, and when reviewing its wrongly classified documents, it was noticed that most of them contained the words "security", "employment", and "work", and thus, were incorrectly categorized in the Security and Employment classes.

Table 2. Confusion matrix for the Bayesian net classifier.

\	**Security**	**Health**	**Economy**	**Employment**	**Education**
Security	12652	199	144	169	72
Health	424	1461	39	52	26
Economy	714	146	1856	403	59
Employment	494	81	280	3961	50
Education	195	55	46	73	1363

On the other hand, modifying the methodology that was expressed in the case of study, in order to use the developed software, would represent a diminution of 70% - 100% of the time employed for data capture, considering that either capture would be done while surveying or, that scanning a paper based questionnaire would take around 7 minutes.

In terms of total precision and complexity of the proposal, as Google's OCR precision and complexity order is not reported in the literature, it is only possible to have an estimation based on assumptions. Considering that in the best cases, OCR algorithms in the literature report having up to a 98% of precision in handwritten texts [22] with a time complexity order of O(n2) [21], while Bayesian net classifiers have reported an order of O(td) for training, and O(cd) for testing where t is the number of training examples, c is the number of classes, and d is the number of attributes [7, 14]. In the worst of the cases, the order for the classifier tends to O(n2), and then the total order for the process is O(2n2).

Finally, considering the error probability on the OCR process of 2%, and an error probability of 14.88% on the classification, the total error probability is given by $P_{TOTAL} = P_{OCR}$ U $P_{BN} = 0.02 + 0.1488 - (0.2 \times 0.1488) = 16.58\%$, where P_{OCR} is the probability of having an error on the OCR process, and P_{BN} is the probability of error on the Bayesian classifier. Next section presents a discussion over the development and the obtained results.

5 Discussion

When analyzing the results that were reported in the previous section, it was possible to highlight some features of the proposal. The first ones corresponding to the change in the work methodology that is required. As it was commented in Section three, at some point of time, the interviewer will need to have Internet access or to be provided with a device for the specific purpose of surveying, representing an additional investment to be done, and leading to ask what happens if information is lost before the

interviewer is able to upload it to the data base? And, how often does the upload process needs to be performed?

The answer to these questionings is directly related to the study in which it is to be applied, depends on the number of interviewers, the amount of surveys to be applied, and the availability of Internet services for the interviewers.

On the other hand, the main advantage that this proposal has against the as-is process is on time consumption. As it was commented in the results section, for the case of study, a saving of a 70% of the invested time was possible. This means that a research study that took 10,400 man-hours in the capture process, may be reduced to 3,120.

Table 3. Confusion matrix for the Bayesian net with 10 folds cross-validation, naïve Bayes with 65% of data on training set, and naïve Bayes with 10 folds cross-validation classifiers.

Bayes net with 10 folds cross-validation					
\	**Security**	**Health**	**Economy**	**Employment**	**Education**
Security	12625	209	150	179	73
Health	430	1452	43	51	26
Economy	723	149	1847	400	59
Employment	493	88	279	3956	50
Education	199	56	50	80	1347
Naïve Bayes with 65% of data as training set					
\	**Security**	**Health**	**Economy**	**Employment**	**Education**
Security	12446	263	173	303	51
Health	645	1226	55	57	19
Economy	995	192	1543	394	54
Employment	591	78	218	3938	41
Education	344	43	51	83	1211
Naïve Bayes with 10 folds cross-validation					
\	**Security**	**Health**	**Economy**	**Employment**	**Education**
Security	12450	267	158	302	59
Health	662	1210	46	57	27
Economy	1005	194	1525	394	60
Employment	592	78	219	3934	43
Education	342	43	49	82	1216

In addition, considering that this specific experiment was done in a period of three weeks, by 65 capturers, using the here proposed capture application, this could be done in the same three weeks by 20 people. This means that the time reduction is translated into an economic saving in logistics and human resources.

It is important also to emphasize that the accuracy of this proposal is determined by at least two elements: the OCR precision, and the classifier training, The OCR outcome depends on the quality of the algorithm, but also on the input (handwritten text is harder to be recognized).

Besides, during the classifier training task, the manual classification may be subjective taking into account the principles that were established in Section three "all responses must fit in a category, and every answer can only be labeled with one class, even if two or more are applicable". Subjectivity in this phase may lead to incorrect training, or inaccurate results during the validation. For instance, consider the following document taken from the case of study: "*falta de trabajo causante de la inseguridad*" which translated to English, is read as "*lack of employment causes insecurity*", and may be manually classified either as an *employment*, or as a *security* matter.

In regards to the Bayesian net classifier which was used, it was possible to contrast it with a Naïve Bayes classifier, and to evaluate both alternatives using a 10 folds cross-validation. The complete results are presented in Table 3. From those results, it is possible to calculate total precision for each alternative, and moreover for each class as it was done in results section for the Bayes net classifier. The comparison of those results is provided in Table 4.

Table 4. Comparison of results for the different classifiers.

\	**Security**	**Health**	**Economy**	**Employment**	**Education**
Bayes net with 65% of data in training set	0.95	0.72	0.58	0.81	0.78
Bayes net with 10 folds cross-validation	0.95	0.72	0.58	0.81	0.77
Naïve Bayes with 65% of data in training set	0.94	0.61	0.48	0.80	0.69
Naïve Bayes with 10 folds cross-validation	0.94	0.60	0.47	0.80	0.70

Using that comparison, it is possible to notice a slight advantage of the Bayes net towards the other alternative, which is translated as a 0.3% benefit on the usage of this method. After reporting results and obtaining insights about them, it is possible to have conclusions from this work, and to establish new objectives. Next section is dedicated to describe those findings inspiring future work on this subject.

6 Conclusions and Future Work

In this paper, the development of a web based application for gathering, classifying, and presenting results from social surveys, using OCR, ML, and IV techniques, was addressed. When considering the specific objectives that were settled in the introduction, it is possible to notice that they were aimed and reached.

The process of gathering data from paper based, and digital questionnaires was covered by developing a tool for either surveying users, or uploading answered surveys, and to extract information from them supported with Google´s OCR. The gathered information was then classified using a Bayes net classifier, which performance was measured and contrasted in terms of precision and complexity order. Finally, the presentation of the classified information was supported with the implementation of a real-time updated module that receives information specific to the domain, and using the InfoViz library, provides graphical outcome representing the results of the elaborated surveys.

Following the idea of having control of the whole process, and giving a more straightforward execution to the user, it was identified that using an OCR API is not the ideal option. Having a tailored implementation of an OCR off-line algorithm supported by digital image processing techniques for segmentation, would allow the evaluation of it in terms of precision and complexity, leading to the correct validation of the entire proposal. Regarding the classifier, in order to increase the precision of this proposal, two lines may be followed as future work: exploring automatic unassisted ML techniques, and considering other assisted techniques, and compare their results with those reported in this paper. From the domain perspective, not all data may be susceptible of being displayed in a tabular way, but could be more illustrative when shown in a geographical manner through information layers on maps. The InfoViz library allows such sort of tasks, but a modification on the questionnaire creation, data gathering and information classification processes will be required.

References

1. Aggarwal, C.C., Zhai, C.: A survey of text classification algorithms. Mining text data, Springer Science & Business Media, Springer, Heidelberg, 163–213 (2012)
2. Althaus, S.L.: Collective preferences in democratic politics: Opinion surveys and the will of the people. Cambridge University Press, New York (2003)
3. Barlett, J.E., Kotrlik, J. W., Higgins, C.C.: Organizational research: Determining appropriate sample size in survey research. Information technology, learning, and performance journal, 19, 43–50 (2001)
4. de Mello, C.A., Lins, R.D.: A comparative study on OCR tools. In: Vision Interface'99 Conference, 224–231 (1999)
5. Downie, N.: Open source charts for your website. http://www.chartjs.org
6. Fekete, J.D., Van Wijk, J.J., Stasko, J.T., North, C.: The value of information visualization. Information visualization, LNCS, 4950, Springer, Berlin, Heidelberg, 1–18 (2008). doi: 10.1007/978-3-540-70956-5_1

7. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine learning, 29, Springer, 131–163 (1997). doi: 10.1023/A:1007465528199
8. Gilbert, D.: The jfreechart class library. Developer Guide, Object Refinery, 7 (2002)
9. Handel, P. W.: U.S. Patent No. 1,915,993. Washington, DC: U.S. Patent and Trademark Office (1933)
10. Ilieva, J., Baron, S., Healey, N.M.: Online surveys in marketing research: Pros and cons. International Journal of Market Research, 44, 361 (2002)
11. Joachims, T., Sebastiani, F.: Guest editors' introduction to the special issue on automated text categorization. Journal of Intelligent Information Systems, 18, 103–105 (2002)
12. Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: A review of classification techniques. Frontiers in artificial intelligence and applications, 160, IOS Press, Netherlands (2007)
13. Langley, P., Iba, W., Thompson, K.: An analysis of Bayesian classifiers. Association for the advancement of artificial intelligence journal, 90, 223–228 (1992)
14. Martınez, A.M., Webb, G.I., Chen, S., Zaidi, N.A.: Scalable learning of Bayesian network classifiers. Journal of Machine Learning Research, 17, 1–35 (2016).
15. Mathers, N., Fox, N.J., Hunn, A.: Surveys and questionnaires. Trent RDSU, United Kingdom (2007)
16. Mori, S., Suen, C. Y., Yamamoto, K.: Historical review of OCR research and development. In: Proceedings of the IEEE, IEEE Press, 80, 1029–1058 (1992). doi: 10.1109/5.156468
17. Ravallion, M., Chen, S.: What can new survey data tell us about recent changes in distribution and poverty? The World Bank Economic Review, Oxford University Press, 11, 357–382 (1997). doi: 10.1093/wber/11.2.357
18. Risch, J.S., Rex, D.B., Dowson, S.T., Walters, T.B., May, R. A., Moon, B.D.: The STARLIGHT information visualization system. In: IEEE conference on Information Visualization, 1997, IEEE Press, 42–49 (1997)
19. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR), ACM, New York, 34, 1–47 (2002)
20. Tauschek, G.: U.S. Patent No. 2,026,330. Washington, DC: U.S. Patent and Trademark Office (1935)
21. Yalniz, I.Z., Manmatha, R.: A fast alignment scheme for automatic OCR evaluation of books. In: International Conference on Document Analysis and Recognition (ICDAR), 2011, IEEE Press, 754–758 (2011)
22. Zhang, H., Liu, C.L.: A lattice-based method for keyword spotting in online Chinese handwriting. International Conference on Document Analysis and Recognition (ICDAR), 2011, IEEE Press, 1064–1068 (2011)

ISSN 1870-4069

Towards the Construction of a Clustering Algorithm with Overlap Directed by Query

Beatriz Beltrán, Darnes Vilariño, David Pinto, Rodolfo Martínez

Benemérita Universidad Autónoma de Puebla,
FCC - Doctorado Ingeniería del Lenguaje y del Conocimiento, Mexico

{bbeltran, darnes, dpinto}@cs.buap.mx, beetho@gmail.com

Abstract. Clustering algorithms are one of the most important element for the data mining and for the automatic learning, also, they keep on being a research topic because algorithms have some limitations. In addition, most of algorithms do not bear overlaps of the groups that they generate, however, in some problems like user profile, clustering algorithms are required for bearing overlaps. Therefore, it pretends to develop an overlapping clustering algorithm using different information retrieval techniques in big data quantities and avoiding computational costs.

Keywords. Clustering algorithms, data recovery, classification models, overlapping groups.

1 Statement of the Problem to be Solved

There is lots of information on internet, due to most of enterprises automate all their processes and local storage problems disappear because the information is in the web, then, with the generation of groups appear the possibility and need to identify groups for data analysis, data managing efficiently, inspect trends, give recommendations, filter queries or having organized information.

Most of the developed clustering algorithms do not allow to overlap between their groups because many problems demand that groups are separated. According to the kind of grouping obtained [1]: It could have the classification as ***exclusive***, this mean, one element belongs to one group and the rest of them not, for example, films could be classify by their content (AA, A, B, B15, C y D), in ***fuzzy***, when an element belongs to all the groups, but with a certain grade of belonging, for example, a range of a million colors and finally, one which are ***overlapped*** where an element could belong to more than one group, for example, the food tastes of people.

The development of a clustering algorithm with overlap is proposed for lots of information volumes, but using an *information retrieval system* which enable to obtain groupings to apply them. For example, in several problems, mainly in the language natural processing area, such as, user grouping, recommendation system like products, films, vacation places, etc., user profiling, stablish themes for documents, organ-

ize documents on the network, trendy news and its analysis, trend of political campaigns, etc. So, the paper is described as follows, first it is described a brief state of the art, in section two, it is given the methodology, in section three it is provided the main contributions and in section four there are some results and validation.

There are many proposal of clustering algorithm with overlap, Star algorithm is showed in [2], and different extensions of the Star algorithm in [9, 10], with data mixed [11], document groupings [12], among others. Exist different groups to mention, but the algorithm Star [2] is taken to illustrate this, the algorithm Star proposes a grouping through a graph not directed weight G= {V, E, *w*}, where V is the set of vertexes which represents documents or in general, elements, E is the set of edges with the weight *w*, corresponds with the resemblance between two documents or elements. For the resemblance, the cosine metrics is used, initially, G is a complete graph, and the grouping leaves a graph of partial maximal coverage, with those edges of resemblance σ, it can get lots of overlapping groups, as not overlapped.

A clustering algorithm with incremental overlap is called Incremental Clustering by Strength Decisions is shown in [3], which uses a heuristic of coverage graphs getting large groups, reducing computational costs, but keeping the incremental structure of the grouping. The corpus used, were pre-processed through eliminating of stop words, lemmatization and the representation used was the Vector Space Model. Cosine was used as a measure of similarity.

The clustering algorithm CLOPE [4] is used for categorical data, being fast, scalable and memory optimization. The similarity measure, proposed, take an intuitive idea as a base to increase the weight proportion of the histogram grouping, using principally this measure due to it is easier and more effective than Jaccard or Dice measures. In the experiments made, the algorithm is very effective, gives interesting groups although a dissimilarity measure intra-groups is not given in specific.

In [5] shows a clustering algorithm of documents, called *Generalized Star (GStar)*, which is a generalization of the Star algorithm, this includes a new concept for Star, allowing a new star form with overlapping groups and defining few groups being easy in implementation and efficient, for the tests, Jaccard was used as a similarity measure, the algorithm is useful for groupings that could be required in the organizing information, surfing, tracking of topics and detecting new topics.

Spatial databases also require clustering algorithms, so an algorithm is necessary to do this task, DBSCAN [6], this has minimum requirements of domain knowledge to determine the input parameters, the discovering of arbitrary form groups. DBSCAN is a clustering algorithm based on density, where the results of the experiments show that it is more effective to discover arbitrary form groups than others algorithm.

In [7] shows the incremental algorithm for overlapping groups, called Incremental Clustering by Strength Decision (ICSD), overlapping and dense groups are getting from a heuristic of coverage graph, reducing the computational time, but keeping the grouping incremental structure. The algorithm builds, as it was mention already, from a set of overlapping and dense groups applying the heuristics, this reach the execution time and do more efficient the managing of the multiple insertions in incremental environments.

There is a hierarchical clustering algorithm, call *Hierarchical Compact Algorithm* and its dynamic version make up a framework [8], these work with dynamic and static data sets. It can obtain different hierarchical agglomerative algorithms specifying a similarity measure inter-groups, one sub-graph β-similarity and another coverage algorithm. It could use for tasks which require dynamic clustering, such as organizing information, document taxonomy and hierarchical topic detection.

2 Research Methodology

Taking into consideration that a typical information retrieval system [13], according to the lots of information given, is required having precise and fast access of itself through one question (query). Manually, perform the retrieval of the information brings as mainly consequence that the most important information is ignored, due to it does not have the necessary precision on the process because one person wants to do fast the tasks, then, he usually ignores this information. Also, in many cases physically is impossible to do this task because there are so much information. For this reason, an information retrieval system pretends solve, by mean of the automatic solution of, given a query (user need) and having much information, the system will recover the most relevant elements for the query (Fig. 1).

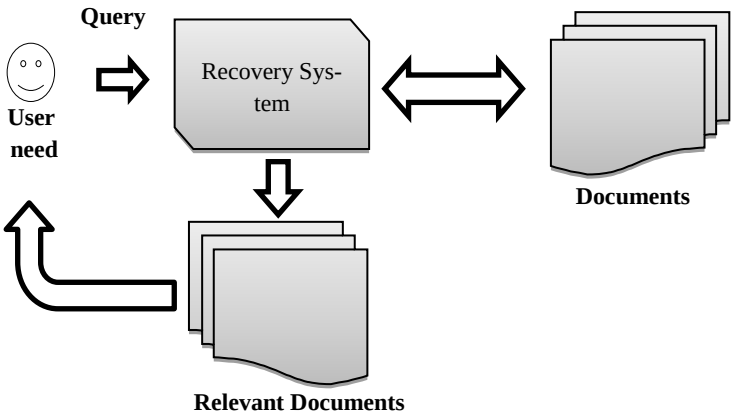

Fig. 1. Typical Information Retrieval System.

In the problem to develop, the query changes to a document to cluster, the documents will be with some presentation and the information retrieval system will propose any similarity measure. Getting the similarity with the documents, the output system will be the similarity with the documents, taking into consideration those one which are over the definite threshold and those one will supply the clustering, this proposed system is showed in the Fig. 2.

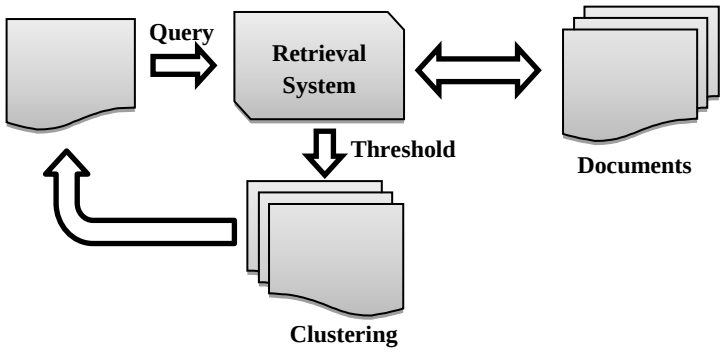

Fig. 2. Proposed System to Generate Clustering.

For all the above reasons, the following methodology is proposed (Fig. 3): first, it is necessary to compile a corpus, for this, it will perform a search of labeled corpus and they should have any overlap in their data. At the same time, a deep analysis of the clustering algorithms proposed is required, from disjointed clustering algorithms like those ones proposed for clustering with overlap, in this analysis, the revision, of the similarity measures used in the different reported cases, will do considering the results obtained.

Design the clustering algorithm with overlap for big information volumes with information retrieval system mechanism. Defining firstly, one or different documents representations, similarity measure to use and define a threshold resemblance.

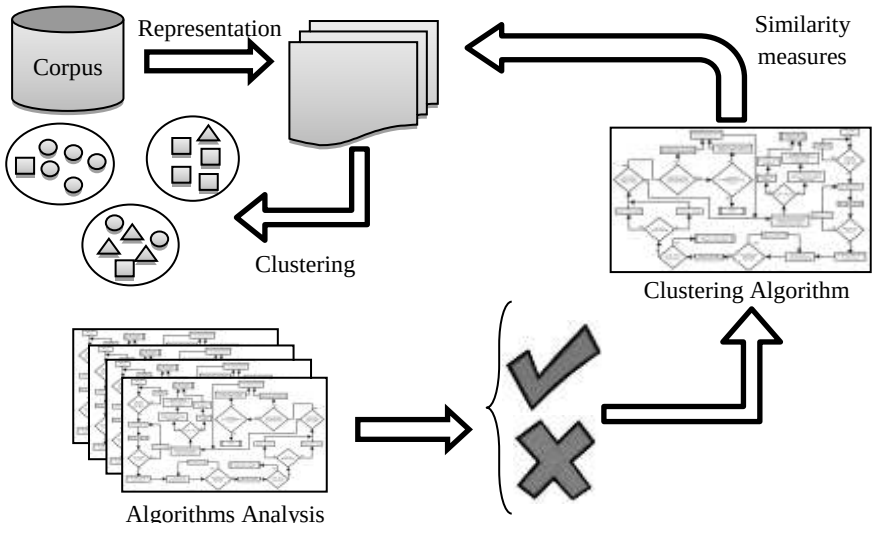

Fig. 3. Methodology proposed.

Implementing the algorithm and do proof to consider with the different representations, similarity measure and thresholds.

Analyze the evaluation ways with the clustering algorithms to check the groups quality and propose one form for evaluating which allow to verify if the algorithm proposed generates competitive groups with the reported algorithms and evaluate the algorithm proposed.

Make a comparison between different algorithm and the proposed to check the quality of the groups generated, considering the execution time facing the big information volumes.

Make an analysis of the results obtained and if it is possible to determine the limitations that could be determine within the algorithm.

In the development of the present research, the main contribution is to use the recovery information system mechanisms as a methodology to model a generalization of the design of one clustering algorithm with overlap.

3 Main Contribution

The contribution is a clustering algorithm with overlapping, using information retrieval system. The first proposal to use an information retrieval system is to use posting list for trigrams and tf-idf. For this algorithm, was applied to corpus of PAN 2017.

The results are significant for cluster algorithms, but this is only a proof and it is necessary more proofs with other corpora, implements some techniques of information retrieval system and it is important to check different threshold.

4 Results and Validity

Some proof was made about the procedure described already, the corpus competence was taken from the PAN[1] 2017 in the author clustering tasks. This consist in 60 problems in 3 languages (English, German and Greek), each problem is made up of 20 texts and have the gold standard. In this paper, the algorithm only was probed with documents in English, but the algorithm could be executed with other languages and only it is presented some results.

The procedure carried out, in the proof made, explain the following algorithm, the input is a query training document and de output is a set of documents with a similarity. The clusters are obtained with those documents with a similarity great or equal to 0.05. this allow the possibility of overlap. The threshold was taken with value great or equal to 0.05, in a way empirical for this analysis. If the threshold is changed, the clusters are modified.

The representation of the training documents and the query was for trigram words, any pre-processing was not made in the documents. The similarity measure was the cosine, the change in this similarity measure will change the groups created.

[1] http://pan.webis.de/

```
SRI Algorithm
```

Input: Query training document
Output: Documents with a threshold greater or equal to 0.05 of similarity.

```
1. PL ← Posting List (training)
2. For each trigram in the query
     Wtq ← (1+log(tfq(trigram)))*log(N/df(trigram))/log
     10
     For each DocId in PL
       Wtd ← (1+log(tf))*log(N/df(trigram))/log 10
       Score[DocId] ← Score[DocId]+(Wtd*Wtq)
     End
3. End
4. For each elem in Score
     Score[elem] ← Score[elem]/TotalDoc
     Write elem, Score[elem]
5. End
```

Different proofs were made, and they show some results. The proof obtained show positive results, the table 1 shows the groups given by the standard gold for the problem 15.

Table 1. Example of group of problem 15.

Groups	Groups of the gold standard
1	document0004 document0011
2	document0001 document0002 document0003 document0006 document0010 document0013 document0016
3	document0005 document0014
4	document0007 document0008 document0009 document0012 document0015

Table 2 shows a query made with the document0001 like a test, the resemblance is presented with a threshold greater than 0.05, the elements of the gold group could be observed that four elements of six are recovered.

Table 2. Similarity between document0001 and the great similarity.

Document	Similarity
document0001	0.695751
document0016	**0.0866128**
document0003	**0.0637452**
document0013	**0.0583805**
document0010	**0.0519872**

Table 3 explains a query made with the document 0007, this show a threshold greater than 0.05, in this case, it could be observed that all documents of the gold standard group, are recovered.

Table 3. Similarity between document0007 and the great similarity.

Document	Similarity
document0007	0.758995
document0015	**0.120644**
document0008	**0.09973**
document0012	**0.0799576**
document0009	**0.0664196**

In the following, the results for the problem 002 are showed, table 4 display the groups supplied by gold standard for this problem.

Table 4. Example of group of problem 002.

Groups	Groups of the gold standard
1	document0005 document0017
2	document0001 document0006 document0009 document0012 document0014 document0015
3	document0003
4	document0004 document0013 document0018
5	document0002 document0010 document0011 document0016 document0019
6	document0008
7	document0007 document0020
8	document0005 document0017

In the table 5 a query was made with the document0001 as a test, the resemblance shows a threshold greater than 0.05, elements of the standard gold could be observed that three of six documents are recovered.

A query was made with document0002 as table 6 shows, the resemblance has a threshold greater than 0.05. elements of the standard gold could be observed that three of five documents are recovered.

Table 5. Similarity between document0001 and the great similarity.

Document	Similarity
document0001	0.73864
document0015	**0.0827088**
document0019	0.0752238
document0014	**0.0694811**
document0012	**0.0589523**
document0016	0.0582658
document0018	0.0574498

Table 6. Similarity between document0002 and the great similarity.

Document	Similarity
document0002	0.824262
document0006	0.0932512
document0010	**0.0862495**
document0019	**0.0827472**
document0016	**0.0709717**
document0008	0.0661357
document0017	0.0653661
document0009	0.0590648
document0020	0.0540831
document0004	0.0519196
document0005	0.051108

These tests confirm that the problem statement made about this proposal could have positive results. This allow in this way, for more corpora in different areas.

In the table 7 is presented the precision and recall for the experiments were showed.

Table 7. Precision and recall for the experiments.

Problem	Query	Precision	Recall	F-measure
15	document0001	1	0.667	0.8
	document0007	1	1	1
2	document0001	0.5	0.6	0.545
	document0002	0.3	0.75	0.429

5 Conclusion and Future Work

In this paper, it was performed some tests with at least one tagged corpus, which allows to obtain results so far satisfactory. It was obtained an average F-measure of 0.69, and this could have some good results, in general.

An information retrieval system was used, using a technique such as *tf-idf*, and making use of the proposed methodology. An algorithm for clustering with overlap-

ping was proposed, with input a query and the output a cluster, using a similarity measure, like cosine.

As future work requires testing with other corpora and testing other thresholds. In addition to making use of other tools of information retrieval systems.

Acknowledgements. We would like to thank VIEP – BUAP and LKE PhD program for supporting this research work.

References

1. Jain, A. and Dubes M.: Algorithms for Clustering Data. Prentice Hall (1988)
2. Aslam, J., Pelekhov, K. and Rus, D.: Static and dynamic information organization with star clusters. In: Proceedings of the seventh international conference on Information and knowledge management ACM, 208–217 (1998)
3. Pérez, J.A., Martínez, T.J., Carrasco, O.J. and Medina, P.J.: A New Incremental Algorithm for Overlapped Clustering. In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications CIARP 2009. Lecture Notes in Computer Science, 5856, 497–504 (2009)
4. Yan, Y., Guan X. and You J: CLOPE: A Fast and Effective Clustering Algorithm from Transactional Data. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, SIGKDD'02 ACM, 682–687 (2002)
5. Pérez, J. A. and Medina P. E.: A Clustering Algorithm Based on Generalized Stars. LNAI, Springer-Verlag, 4571, 248–262 (2007)
6. Ester, M., Kriegel, H. P., Sander, J. and Xu, XS.: A Density - Based Algorithm for Discovering Cluster. In: Proceedings KDD-96 AAAI, 226–231 (1996)
7. Pérez, A. S., Martínez, J. F., Carrasco, O. J. and Medina, P. J.: A New Incremental Algorithm ios for Overlapped Clustering. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 14th Iberoamerican Conference on Pattern Recognition, 5856, 497–504 (2009)
8. Gil, G.R., Badía, C. J. and Pons, P. A.: Dynamic Hierarchical Compact Clustering Algorithm. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 10th Iberoamerican Conference on Pattern Recognition, 3773, 302–310 (2005)
9. Gil-García, R.J., Badía-Contelles, J.M., Pons-Porrata, A.: Extended star clustering algorithm. In: Proceedings of the 8th Iberoamerican Congress on Pattern Recognition (CIARP2003), LNCS 2905, 480–487 (2003)
10. Pérez, S.A. and Medina, P.J.: A clustering algorithm based on generalized stars. In: Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2007), LNAI 4571, 248–262 (2007)
11. Pons, P.A., Ruiz, S.J., Berlanga, L.R. and Santiesteban, A.Y.: Un algoritmo incremental para la obtención de cubrimientos con datos mezclados. Reconocimiento de Patrones Avances y Perspectivas Research on Computing Science, 405–416 (2002)
12. Zamir, O. and Etziony, O.: Web document clustering: A feasibility demonstration. In: Proceedings of the 21st Annual International ACM SIGIR Conference, 46–54 (1998)
13. Manning, C.D., Raghavan, P. and Schütze, H.: An Introduction to Information Retrieval. Cambridge University Press (2009)

ISSN 1870-4069

A Proposal of Lexical Resources' Development for Ontological Learning in the Domain of Speech Disorders

Stephanie Vázquez, María Somodevilla, Ivo Pineda, Concepción Pérez de Celis

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, Puebla, Mexico

{stephanie.vazquez, mariajsomodevilla}@gmail.com,
{ivopinedatorres, mpcelish}@gmail.com

Abstract. Speech disorders in children are a condition that could reduce the opportunity to access education, health care and in the future could mean a worse socioeconomic outcome. Therefore, early diagnosis and timely therapy is really important to reduce their impact in later stages of life. This paper presents a method for the gathering of data for a corpus related to Speech Disorders in children; such corpus will serve as the base to generate a semi-automatic ontology intended as a tool for therapists to help in the diagnosis and shape up of a therapy strategy.

Keywords. Speech disorders, corpus building, crawling, dictionary building, semi-automatic ontology creation.

1 Introduction

A speech disorder is the difficulty to produce or to create the specific speech sounds to communicate. These disorders can range from simple sound substitutions to disability for understanding or using the language (motor-oral mechanism) for the speech. Causes could be as diverse as hearing loss, neurological disorders, brain injury, intellectual disability, or physical impairments as cleft lip [1].

According to Global Disability Rights 7.5% of the population in Mexico has some disability (about 9.17 million people) and 4.87% of people with disability has some type of speech disorder (0.45 million people). In kids and young people the speech disabilities are in some cases twice or four times higher than in adults [2]. The majority of people with disabilities do not have equal access to health care, education, and employment opportunities, do not receive the disability-related services that they require, and experience exclusion from everyday life activities, furthermore a disability is a development issue: evidence shows that persons with disabilities experience worse socioeconomic outcomes and poverty than persons without disabilities [3].

The importance of the early detection and diagnosis of a speech disorder abides in the social, economic and educative impact that such disorders have in the life of infants. Technology is used in order to assist in the process of diagnosis and treatment

of some speech disorders in children. ICT (Information and Communication Technologies) are helpful in almost every step to identify and provide a treatment for a speech disorder.

To deal with the problem of manipulating and organizing a big amount of data such as speech disorders' information the use of ontologies can be resorted to. An ontology provides through Semantic Web -an evolving extension of the World Wide Web- the semantics of information and services so that the Web can understand and satisfy requests for content made by people and machines [4]. Ontologies give an unambiguous and well defined structure for a clear and accurate representation of the data concerning a particular domain, in this case speech disorders, and thus, becoming a tool for diagnosis. Ontologies are made up of two main components: classes and relationships (See Fig. 1).

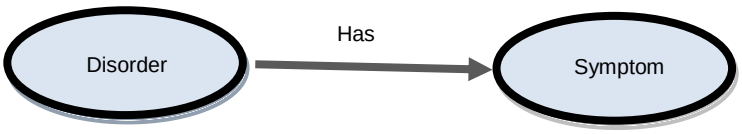

Fig. 1. Simple representation of the two main components in an ontology: classes and relationships.

Is proposed an ontology to organize and to look up the information relative to speech disorders such as different disorders, characteristics of each disorder, therapy plans, taxonomy of the speech disorders, and some other helpful information for the therapist and patient, as well as the relationships between all of them. One of the earlier steps in the development of an ontology is the conformation of a Corpus, in this case of documents relatives to the domain of speech disorders.

A Corpus is a large collection of texts. It is a body of written or spoken material upon which a linguistic analysis is based. The corpus may be composed of written language, spoken language or both. Spoken corpus is usually in the form of audio recordings. A corpus may be open or closed. An open corpus is one which does not claim to contain all data from a specific area while a closed corpus does claim to contain all or nearly all data from a particular field. Computer-processable corpora allow linguists to adopt the principle of total accountability, retrieving all the occurrences of a particular word or structure for inspection or randomly selected samples. Corpus analysis provide lexical information, morphosyntactic information, semantic information and pragmatic information [5].

This document is organized as follows: section 2 presents the state of the art through the discussion of some works related to the subject of the present work. Section 3 exposes the model proposed by the authors to build a corpus as a data source for the future ontology; subsections detail the construction of a very important component in this model: the *dictionary*. Section 4 presents the data obtained when testing the corpus with several algorithms and the resulting extended dictionary. Finally, in Section 5 the conclusions of applying the proposed model are outlined followed by the references.

2 State of the Art

Within the field of speech and language several works that use Information and Communication Technologies (ICT) have been conducted, focusing on some ailments such as dysphagia [6], on the automatic classification of the quality of pronunciation when treating disorders such as dyslalia or dysarthria [7], or an expert system for the initial evaluation of children with possible speech disorders [8]. A so-called ecosystem of smart ICTs that include electronic medical record management, standardized vocabularies, a knowledge database, ontologies for concepts within the domain of speech and language, and expert systems focused on supporting speech and language pathologists, doctors, students, patients, and their relatives can also be found [9]. There are also tools for the formation of professionals in the field of speech disorders based on ontologies and e-learning, which support future language therapists in their training process, as well as in their development of practical abilities [10]. Regarding language therapies, a mobile app that integrates therapy activities for children and that uses colloquial language, as well as games from the state of Chiapas, has been developed [11]. There even is a robust ontology that covers several aspects of speech and language therapies, with key concepts such as initial evaluation and patient profile, conducted tests, doctors and therapists catalog, list of disorders, speech and language fields, therapy and tracking plans and exercises, among others, that uses OpenEHR ontologies and constructs [12].

Regarding the semi-automatic creation of taxonomies for a given domain, several methods have been proposed that use techniques as diverse as formal Horn concepts and clauses analysis through logical inference validation [13], hierarchical clustering of documents based on sets of frequent concepts validated through prototype implementation [14], or a generalized algorithm of association rules that detects relationships between concepts and that detects the proper abstraction level for relationships definition [15].

Relevant to the building of corpus the main techniques have not varied a lot, and texts in a corpus need to be in electronic form. Thus, the fastest way to build a corpus is gathering data that is already digitalized or relying mainly in transcript into electronic form the audios, or documents [16].

In the present work, a method to gather information for the corpus building is proposed. This method also has the flexibility to feedback itself; once the initial dictionary is defined this can be updated with the extended dictionary obtained after completing the several steps into the method.

3 Information Retrieval Model for the Definition of Lexical Resources

In order to build the Corpus, it is necessary to gather a big amount of documents relevant to speech disorders through a Web Crawler. This crawler uses a predefined dictionary with some of the terms relevant to the domain. Once a representative amount

of those documents is obtained, they need to be pre-processed in several steps to clean up and standardize the data thought algorithms like normalization and stemming. Once the data is clean and ready to retrieve information some algorithms like word ranking and n-grams are applied to extend the original dictionary relevant to the domain of speech disorders. In this section each step in the conformation of a corpus will be explained.

The several steps in the task of building corpus and processing it can be seen in a diagram in Fig. 2.

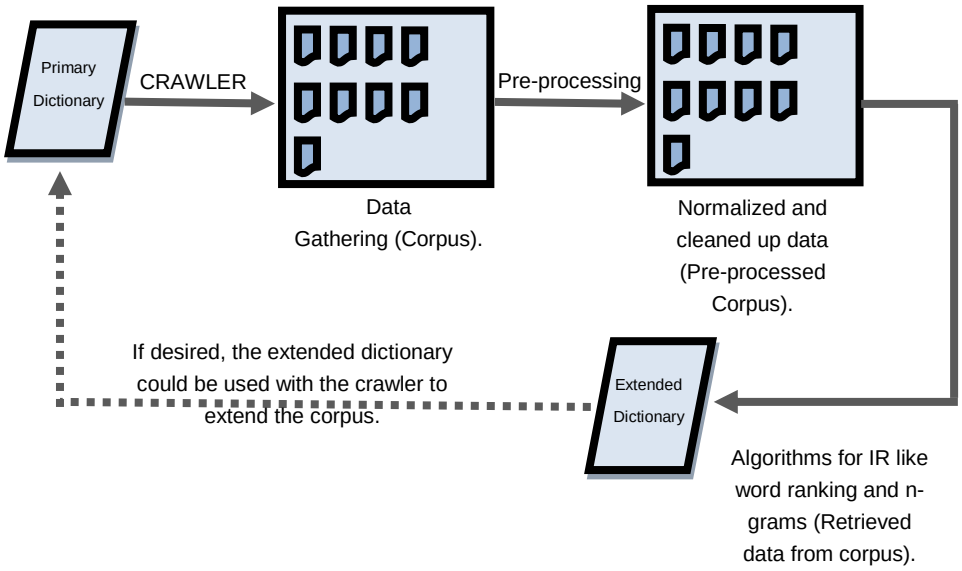

Fig. 2. Diagram of the steps to build and process a corpus.

3.1 **Corpus Creation**

The building of a corpus is divided into two stages: design and implementation. A good practice in the stage of design is to define what would ideally the corpus will have, in terms of the amount and the type of language, and then the parameters could be adjusted as the building goes along, keeping a careful record of what is in the corpus, so it can be added and amended later, and if others use the corpus they know what is in it [16].

In order to build a corpus there are a number of factors which need to be taken into consideration. These include size, balance and representativeness. The size of the corpus depends very much on the type of questions that are going to be asked of it.

The sample documents in our corpus would need to be balanced. Getting this balance right is not an exact science and there are no reliable ways of determining whether a corpus is truly balanced. One approach to achieving balance is to use an existing corpus as a model; research has suggested that samples of 2,000 to 5,000 words are sufficient.

A corpus can be said to be representative if the findings from that corpus are generalizable to language or a particular aspect of language as a whole. The notion of 'saturation' can be used. Saturation (at the lexical level) can be tested for by taking a corpus and dividing it into equal sections in terms of number of words. If another section of the same size is now added, the number of new items in the new section should be approximately the same as in the other sections [17].

The main tool to gather the information to build a corpus is a Web crawler. A crawler can be defined as an Internet *bot* that browses the World Wide Web, typically with the purpose of Web indexing. This crawler is fed with some initial *seed* pages to start its task. At their core is an element of recursion. They must retrieve page contents from an URL, examine that page for another URL, and retrieve that page, ad infinitum [18]. To find documents relevant to the domain, and not just a list of links and random data contained into the seed page, it is necessary to establish a primary dictionary at the beginning of the crawling.

3.2 **Dictionary Creation**

This dictionary is made of some of the more significative words into the domain. A simple way to identify these words is to take the domain taxonomy as a base to gather such list of words. The figure3 shows the taxonomy of speech disorders proposed by the DSM-5 manual of Mental Disorders [19].

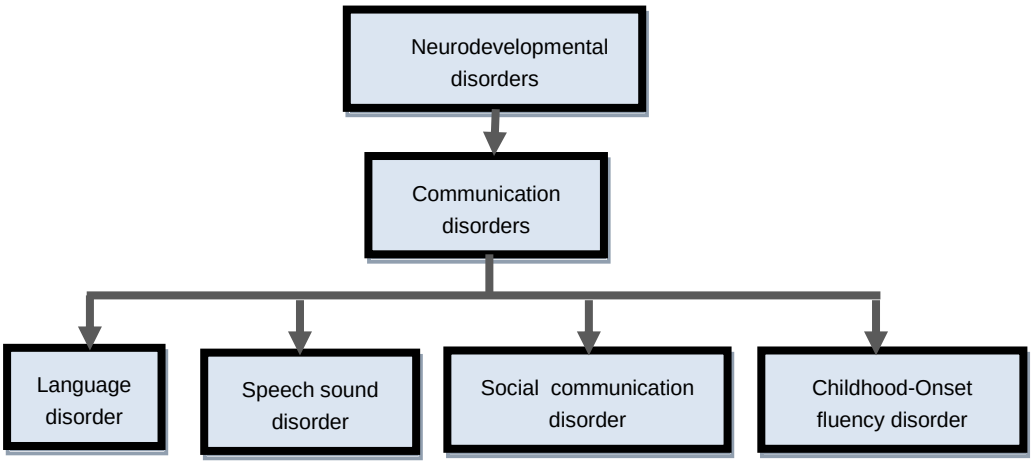

Fig. 3. Hierarchical Taxonomy of Speech Disorders according to DSM-5 manual.

Then the building of the primary dictionary to focus the results of the crawler can be started. As this taxonomy is a small one, the size of the dictionary using some other terms related to the ones included in our taxonomy could be increased. There are some other classifications for speech disorders that include specific names for each kind of *speech sound disorder*. In Fig. 4 another example of speech disorders classification can be seen.

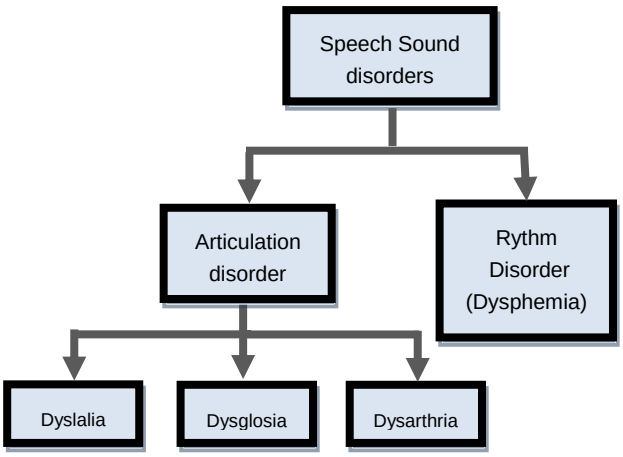

Fig. 4. Additional classification for speech sound disorders.

Since the ontology is focused only in the diagnosis and therapy of *speech sound* and *fluency* disorders are used just the relevant terms relative to such disorders. The table 1 shows the very first version of the primary dictionary.

Table 1. List of terms from the primary dictionary.

No.	Term(s)		Term(s)		Term(s)
1	Speech	6	Dysphemia	11	Rhythm disorder
2	Disorder	7	Speech sound disorder	12	Therapy
3	Dyslalia	8	Childhood-onset fluency disorder	13	Speech therapy
4	Dysglosia	9	Communication disorder	14	Logopedic therapy
5	Dysarthria	10	Articulation disorder	15	Speech development

Starting with the terms directly obtained from the branches of the taxonomies that are related with the domain, terms as *communication disorders, speech sound disorders, childhood-onset fluency disorder, articulation disorders, rhythm disorder, dyslalia, dysglosia, dysarthria, dysphemia* and every single significative word in those terms goes to the dictionary. As the ontology will also contain data about the therapies applied to the previously listed disorders it's also desirable to include related terms like *therapy, speech therapy, logopedic therapy* (the study and treatment of speech defects) and *speech development.*

The next step is the use of this dictionary to gather the corpus for the ontology using a web crawler written in Python language with the help of libraries *HTMLparser, urlopen* and *BeautifulSoup* [20]. Using some Web pages relevant to the domain of speech disorders (like *www.asha.org, medlineplus.com*, etc.) the traversing of those sites is started in search of each term of the dictionary once at a time and retrieving the data in each site visited, store that data in a file and then storing the links to and

visiting internal pages into the *seed page* provided like parameter to the crawler. An additional parameter for the crawler could be a maximum number of pages to visit in search of the term. After retrieving relevant data for all the primary dictionary terms the first version of our corpus is finished, but the processing of the corpus is not done.

3.3 **Data Preprocessing**

Pre-processing the data is the next step. This is done through several algorithms that normalize the texts contained in the corpus. Algorithms for removing escape characters, Unicode characters, punctuation marks, stop words, converting to plain text and capitalization are very useful to clean-up the data before being analyzed [21][22]. Again, with some Python routines are performed the algorithms to clean the data. Once all the data gathered into the corpus is normalized the next step in the process can be done.

In this step, information retrieval algorithms are implemented. Algorithms like word frequency and stemming are used [22]. After this last step a new list of terms for the extended dictionary is obtained. The more frequent terms found into the corpus are taken and is made a comparison with the primary dictionary terms. In the following section this comparison and some additional data about the corpus and the data included into it are presented.

4 Testing

As a result of the Web crawling using as seeds the terms from the first version of the dictionary, as shown in Table 1, an amount of documents relevant to the subject of speech disorders was obtained. Some data about the corpus is now presented in Table 2.

Table 2. Some outline data from the corpus.

Number of initial terms used in the gathering of documents.	15
Number of offline documents added to the corpus.	25
Number of documents obtained at the end of web crawling.	395
Corpus size of plain text in bytes.	3,151,819

After applying the pre-preprocessing described in the previous section and the information retrieval algorithms, the terms shown in Table 3 were found to be the most frequent.

The proposed primary dictionary also included composed terms but in this initial analysis of the corpus just single terms frequency is searched. The original single terms proposed in the primary dictionary can be compared against the single terms found to be the most frequent in the corpus (See Table 4).

Table 3. 15 most frequent terms in corpus.

No.	Term	Frequency
1	Speech	4,036
2	Disorder	2,798
3	Child	2,369
4	Language	1,695
5	Health	1,332
6	Information	1,180
7	Help	963
8	Therapy	949
9	Sound	809
10	Communication	772
11	Research	742
12	Services	695
13	Words	694
14	Development	651
15	Medical	640

Table 4. Comparison of single proposed terms vs single most frequent terms.

No.	Single Term (proposed)	Single term (most frequent)
1	Speech	Speech
2	Disorder	Disorder
3	Dyslalia	Child
4	Dysglosia	Language
5	Dysarthria	Health
6	Dysphemia	Information
7	Therapy	Help

Only the terms *speech* and *disorder* are kept in both lists. Just the first 7 single most frequent terms are used because in the original proposed terms there are just 7 single terms. The rest of the terms that do not appear in the top 7 frequent single terms (*dyslalia, dysglosia, dysarthria, dysphemia* and *therapy*) are listed with their frequency in the Table 5.

Table 5. Frequency in corpus of the rest of proposal single terms.

.

No.	Single Term (proposed)	Frequency in corpus
1	Dyslalia	81
2	Dysglosia	4
3	Dysarthria	437
4	Dysphemia	46
5	Therapy	967

Observing this data from word frequency, not all of the proposed terms in the primary dictionary are equally relevant to the domain of knowledge. Therefore, the web crawler can be fed with the most frequent terms obtained from the corpus and thus, gather more relevant documents.

Another way to complement the corpus is to include synonyms to the original proposed terms. A vast list of terms was found to be included in such list, some of them are listed in the Table 6.

Table 6. Synonyms for some of the original proposed terms.

No.	Original proposed term(s)	Synonyms
1	Speech	Conversation, locution, expression, language, articulation.
2	Disorder	Irregularity, impairment, deficit.
3	Dyslalia	Dysphasia.
4	Dysarthria	Aphasia.
5	Dysphemia, Childhood-onset fluency disorder, Rhythm disorder.	Stammering, stuttering.
6	Speech sound disorder, Communication disorder, Articulation disorder.	Speech impairment, speech impediment, speech defect, delayed speech, speech deficit, speech deficiency, speech disturbance, misarticulation, phonological disorder, phonological delay, phonological impairment, verbal disorder.
7	Therapy.	Treatment, Care.
8	Speech therapy, Logopedic therapy.	Language therapy, Articulation therapy, Speech treatment.
9	Speech development	Speech progress, Speech improvement, Speech maturation, Speech progression.

Applying again the steps of crawling, pre-processing and IR algorithms more documents were added to the corpus and a new list of the most frequent terms is obtained. The Table 7 compares the more frequent terms in the corpus from the last step presented in Table 3 vs. the most frequent terms in the corpus after gathering documents using the synonyms as seeds for crawling.

The 15 most frequent terms obtained after this expansion in the dictionary resulted to be the same as the ones obtained in the previous step non-using synonyms, just varying the order of appearance in the list. Terms as *child* and *language* resulted to be more frequent when synonyms were used as seeds than in the first term frequency list in Table 3.

Table 7. 15 most frequent terms in corpus.

	Primary Dictionary Terms		Extended with Synonyms Dictionary Terms	
No.	**Term**	**Frequency**	**Term**	**Frequency**
1	Speech	4,036	Speech	9125
2	Disorder	2,798	Child	5877
3	Child	2,369	Language	5165
4	Language	1,695	Disorder	4792
5	Health	1,332	Sound	2968
6	Information	1,180	Word	2790
7	Help	963	Health	2786
8	Therapy	949	Information	2697
9	Sound	809	Therapy	2695
10	Communication	772	Help	2081
11	Research	742	Service	1939
12	Service	695	Communication	1687
13	Word	694	Development	1485
14	Development	651	Research	1476
15	Medical	640	Medical	1261

5 Conclusions

The corpus building process proposed for a certain knowledge domain starts with a list of proposed terms followed by a crawling script execution to gather relevant documents. Afterwards, normalizing and IR algorithms were applied to the documents in the corpus in order to include the resulting list of terms into the dictionary; the crawler can be fed again with the new dictionary. Ongoing work consists on the application of word ranking and n-grams algorithms in order to improve the list of terms into the dictionary. Besides, work has been doing in expanding with hyponyms and hyperonyms in the list of terms; this current task allows adding an additional semantic level to the process and it to be able to gather even more relevant documents for the corpus.

Acknowledgements. We would like to thank to the Vicerrectoría de Investigación y Estudios de Posgrado (VIEP) from the Benemérita Universidad Autónoma de Puebla for supporting this work through the project Model of Teaching-Learning Process applying Ontological Engineering.

References

1. NICHCY: Trastornos del habla o lenguaje, 285, 1–4 (2010)
2. Disability in Mexico | Global Disability RightsNow!, http://www.globaldisabilityrightsnow.org/infographics/disability-mexico
3. WHO (World Health Organization): World report on disability 2011. Am. J. Phys. Med. Rehabil. Assoc. Acad. Physiatr (2011). doi:10.1136/ip.2007.018143
4. Loudon, K.: Developing Large Web Applications. O'Reilly Media, California (2010)
5. Robin: What is Corpus?, http://language.worldofcomputing.net/linguistics/introduction/what-is-corpus.html
6. Sharma, S., Ward, E.C., Burns, C., Theodoros, D., Russell, T.: Assessing dysphagia via telerehabilitation: Patient perceptions and satisfaction. Int. J. Speech, Lang. Pathol, 15, 176-183 (2013). doi:10.3109/17549507.2012.689333
7. Schipor, O.A., Pentiuc, S.G., Schipor, M.D.: Automatic assessment of pronunciation quality of children within assisted speech therapy. Elektron, ir Elektrotechnika, 122, 15–18 (2012). doi:10.5755/j01.eee.122.6.1813
8. Martín Ruiz, M.L., Valero Duboy, M.Á., Torcal Loriente, C., Pau de la Cruz, I.: Evaluating a web-based clinical decision support system for language disorders screening in a nursery school. J. Med. Internet Res, 16, e139 (2014). doi:10.2196/jmir.3263
9. Robles-Bykbaev, V., López-Nores, M., Pazos-Arias, J., Quisi-Peralta, D., García-Duque, J.: An Ecosystem of Intelligent ICT Tools for Speech-Language Therapy Based on a Formal Knowledge Model. Stud. Health Technol. Inform., 216, 50–54 (2015). doi:10.3233/978-1-61499-564-7-50
10. Chuchuca-Mendez, F., Robles-Bykbaev, V., Vanegas-Peralta, P., Lucero-Saldana, J., Lopez-Nores, M., Pazos-Arias, J.: An educative environment based on ontologies and e-learning for training on design of speech-language therapy plans for children with disabilities and communication disorders. CACIDI 2016 - Congr. Aergentino Ciencias la Inform. y Desarro. Investigación (2016). doi:10.1109/CACIDI.2016.7785987
11. Ilda, R., Torres, B., López, I.V., Luis, J., Suarez, D.: Aplicación Móvil para la Adquisición de Lenguaje En Niños Con Trastorno De Habla. 40–56 (2016)
12. Kalra, D., Beale, T., Heard, S.: The OpenEHR foundation. http://www.openehr.org/home
13. Haav, H.M.: A Semi-automatic Method to Ontology Design by Using FCA. CLA, 13–24 (2004)
14. Braga, F., Ebecken, N.: A semi-automatic method for extracting a taxonomy for nuclear knowledge using hierarchical document clustering based on concept sets Fabiane Braga. Int. J. Nucl. Knowl. Manag., 6, 155–169 (2013). doi:10.1504/IJNKM.2013.054496
15. Maedche, A., Staab, S.: Semi-automatic engineering of ontologies from text. Proc. 12th Int. Conf. Softw. Eng. Knowl. Eng., 231–239 (2000)
16. Wynne, M.: Developing Linguistic a Guide to Good Practice Corpora (2005)

17. Evans, D.: Corpus building and investigation for the Humanities : An on-line information pack about corpus investigation techniques for the Humanities. Linguistics. 15–16 (2004). doi:10.1109/SocialCom.2010.106
18. Mitchell, R.: Web scraping with Python: collecting data from the modern web. O'Reilly Media, Inc (2015)
19. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders (2013)
20. Python Software Foundation. https://www.python.org/
21. Caio Miyashiro: Text Mining and Natural Language Processing - Preprocessing. http://rstudio-pubs-static.s3.amazonaws.com/67435_ca0769f0dbbb4fc4bda5e4535e21fb54.html
22. Zhu, X.: Common Preprocessing Steps. CS769 Spring 2010 Adv. Nat. Lang. Process, 1–3 (2010)

ISSN 1870-4069

JScheduling: A Graphical Interface for Applying a Process Scheduling Algorithm

Adriana Hernández-Beristain[1], Erika Annabel Martínez-Mirón[1], Mariano Larios-Gómez[1], Javier Caldera-Miguel[2], Luis Angel Zamarripa-Almazan[1]

[1] Benemérita Universidad Autónoma de Puebla, Puebla, Mexico

[2] Universidad Politécnica de Puebla, Puebla, Mexico

adriana_beristain@hotmail.com, mlarios777@gmail.com, erika.a.mtzm@gmail.com, javiercmiguel@hotmail.com

Abstract. Among the tasks that scheduling algorithms perform are the provision of time to each node and processes management, either in a task queue or in a list. It is possible, via a command line terminal, to track the performance of these algorithms. Nevertheless, a visual environment that facilitates this tracking would be very helpful. This work describes the design and implementation of a didactic graphical interface for a distributed embedded software, which allows the visual representation of a process scheduling algorithm in a virtual mobile distributed system.

Keywords. Distributed system, mobile device, embedded software.

1 Introduction

This work describes the implementation of a graphical interface, named JScheduling, for an embedded software; this software shows the use of a process scheduling algorithm that allows to a supercomputer the allocation of its resources optimally among the different nodes connected to it.

A well-known process scheduling algorithm is Simple Linux Utility for Resource Management (SLURM) [1], which is open code and is implemented in many supercomputers based in Linux. This scheduler gives the nodes a time to execute their tasks, manages the processes launched by each node and considers a queue of pending tasks for accessing resources. Besides, SLURM optimizes the nodes' allocation with an algorithm centered in the Hilbert curve [2].

Another process scheduling algorithm, which will be named as "fan", gives the nodes the same amount of time to execute their tasks and, after reaching a consensus, decide the access to the resources.

Sometimes, the understanding of these algorithms can represent a difficulty due to the associated abstraction. So, the implementation of JScheduling to represent a mobile

distributed system, where the user can manage the nodes in a simple, easy and transparent way, as well as visualize how each node executes its processes, becomes a very useful and didactic tool.

The following sections describe the implementation of JScheduling in order to represent, configure, and communicate the nodes, as well as how the fan process scheduling algorithm is used in order to give access to the resources to each node.

2 Development of JScheduling

During the requirement analysis and specification phase, the functionalities identified were: a) the node network creation (being the nodes the mobile devices), and b) the obtaining of the required information to implement a process scheduler under a virtual environment. The following subsections describe how these functionalities were achieved.

2.1 Node Representation

In order to represent a virtual mobile device some considerations were taken into account: 1) a simple design and, 2) to insert as many mobile devices as possible. The first was achieved by using a simple rectangle to represent a node, straight lines for its connections (see Figure 1) and the use of the right click of the mouse for editing options (more details in Section 2.2). It is important to mention that on the back-end of the software, a single adjacency list with four elements (transmitter node, receiver node, bridge node and router node), was used to store the connections. The latter was accomplished combining the computer's date and time and generating a unique identifier for each node.

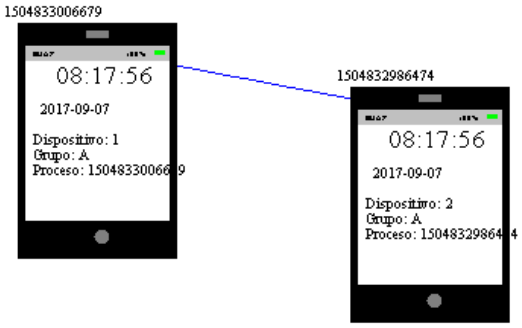

Fig. 1. Representation of two nodes with a single connection.

In addition, for representing the mobile devices and their respective resource requests, Petri networks [3] were used because of their graph design. A graph is defined as a triplet G = <P, R, A>, where P is the set of processes used by the scheduling algorithm, R is the set of resources used in a distributed environment and A is the set of

edges that relate one to another process [4]. Figure 2 depicts an example where processes p1 and p2 request the resources, but it is the process p1 that gets the critical section.

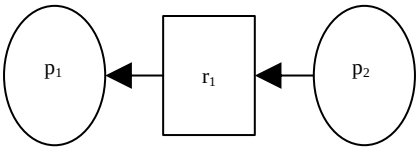

Fig. 2. Example of Petri nets notation.

2.2 Node Configuration

For the creation and manipulation of each node on the workspace, a contextual control menu can be used (right click) to display a set of possible actions: a) New SmartPhone; b) Dragging; c) Connection; d) Individual elimination; e) Overall elimination; f) Intercommunication - JNI y g) Cancellation. Figure 3 shows the menu highlighting the Java Native Interface (JNI) function, that allows the native calls for the communication between the embedded software and the underneath operative system (OS).

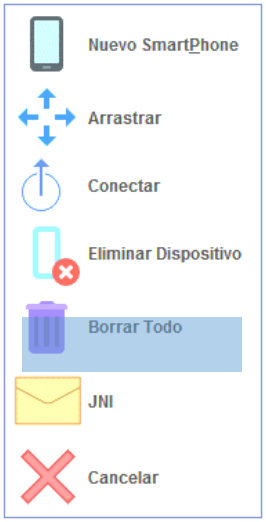

Fig. 3. Contextual control menu.

2.3 Communication between Nodes

The core functionality of JScheduling is the communication between the virtual mobile devices through an information flow between two layers, the graphic layer and the low level layer. For this reason, a multiplatform with JNI and the supercomputer's OS (Minix) was implemented.

Minix is a modular operative system in which services are considered as processes, in comparison to other operative systems, where services are treated as simple calls to the system [5]. In Minix, any kind of processes can communicate between them through primitives to exchange information by means of messages.

In JScheduling, the information exchanged between the nodes includes the node's identifier, the execution time (start, final) and the quantum assigned by the system to the process. The code lines shown in Figure 4 correspond to the function that requests for time in the process planning algorithm and assign the respective ID to the node.

```
  #include <jni.h>
  #include <stdio.h>
  #include <string.h>
  #include <time.h>
  JNIEXPORT void JNICALL Java_DelayTime
    (JNIEnv * env, jobject jobj,jstring i){
   int ID;
   time_t rawtime;
    struct tm * timeinfo;
    time ( &rawtime );
    timeinfo = localtime ( &rawtime );
    const char *str = (*env)->GetStringUTFChars(env, i,
0);
    printf("Dispositivo: %d  Grupo: A  Proceso: %s  Date:
%s  \n",
    ID++, str,asctime (timeinfo));
  }
```

Fig. 4. ID generation for each node in the low level layer.

3 Functioning of JScheduling

For representing the process scheduling, a driver capable of using real resources was implemented as a DLL file. When the JNI option is selected in the contextual menu, the driver requests the ID processes of the connected nodes from the supercomputer working on real time. Then a command line terminal is used to verify that the exchange of messages between the connected nodes was correct (see Figure 5).

Fig. 5. Exchange of messages between emitter (Dispositivo 2) and receptor (Dispositivo 3) nodes.

3.1 Characteristics of the Evaluation Environment

Because a personal computer did not have enough resources for testing the creation of more than five mobile devices, there was the need to use a supercomputer, whose characteristics are described next: The Cuetlaxcoapan supercomputer of the LNS is composed of a standard calculation cluster with Intel Xeon processors and a cluster with Intel Xeon Phi Knights Landing processors with 228 Thin calculation nodes (5472 total cores). Each node contains 2 Intel Xeon E5-2680 v3 processors (Haswell) at 2.5 GHz, 12 cores per processor / 24 total cores, 128 GB of DDR4 memory at 2133 MHz, 2 Gigabit Ethernet network interfaces and an InfiniBand FDR 56 Gbps network interface. Storage of 1.2 PB of total space for disk storage. Finally, there is a Gigabit Ethernet network for managing the hardware of the supercomputer and the provisioning of software to the nodes.

4 Results

The implementation of an interface that allows to represent any mobile distributed system was achieved. The mobile devices are introduced as single rectangles, the straight lines denote the connection between them, and the information inside each node correspond to the resources requested to the system. The interface is centered in graphs, edges and relations between the mobile devices.

Besides the graphical representation of the mobile distributed network, it is possible to visualize the functioning of a process planning algorithm (fan) by means of the information displayed in each mobile device representation. Figure 6 shows how the graphical interface and the process planning algorithm are working together.

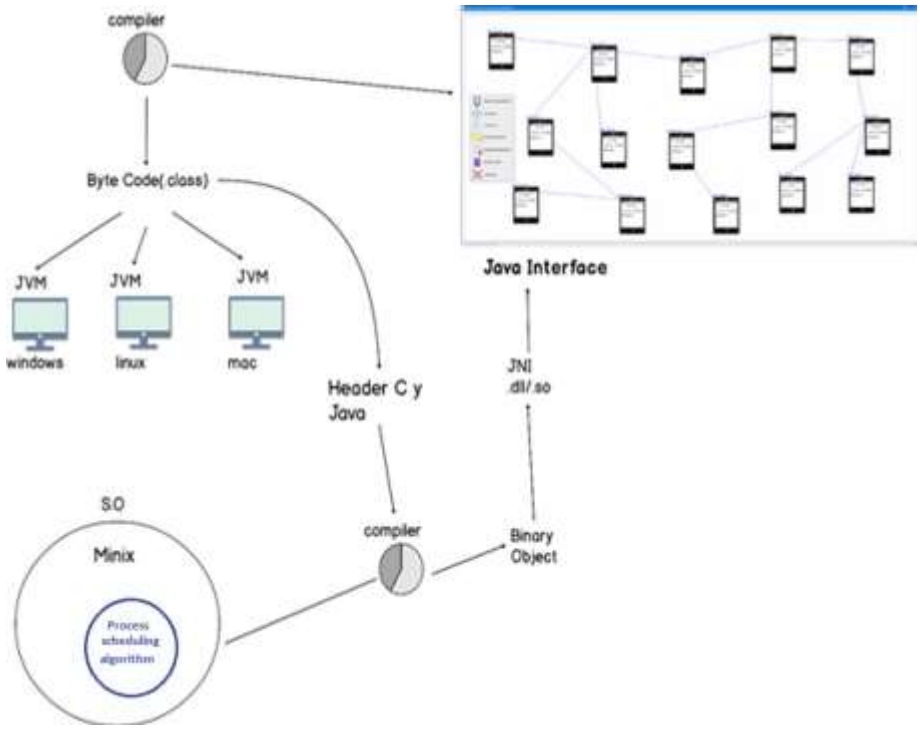

Fig. 6. The graphical interface and the process scheduling algorithm.

5 Conclusions and Future Work

The implemented graphical interface allows the visual representation of a mobile distributed system in a simple, easy and transparent way. At the same time, it is possible to visualize how a fan process scheduling algorithm is executed through the exchange of information between the nodes by means of labels inserted in each node.

So far, just one process can be executed by each node, but it is planned to modify the functions to allow the generation of more processes. Also, the use of other process scheduling algorithms is considered.

Acknowledgments. The authors thankfully acknowledge the computer resources, technical expertise and support provided by the Laboratorio Nacional de Supercómputo del Sureste de México, CONACYT national laboratories network.

References

1. Trinitis, C., Weidendorfer, J.: Co-Scheduling of HPC Applications. 71–73 (2017)

2. Costa, L., Oliveir. P.: An elistist genetic algorithm for multiobjective optimization. 309–310 (2003)
3. Tanenbaum, A. S., Wetherall, D.J.: Computer networks. 232–237 (214)
4. Lee, J. S.: A Petri net design of command filters for semiautonomous mobile sensor networks. IEEE Transactions on Industrial Electronics, 55(4), 1835–1841 (2008)
5. Herder, J. N., Bos, H., Gras, B., Homburg, P., Tanenbaum, A. S.: MINIX 3: A highly reliable, self-repairing operating system. ACM SIGOPS Operating Systems Review, 40(3), 80–89 (2006)

ISSN 1870-4069

Modelo ontológico para representar información sobre la práctica profesional en una institución educativa

Juan Carlos Flores, Mireya Tovar, Ana-Patricia Cervantes

Benémerita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, Puebla, Mexico

{fmjuancarlos}@gmail.com, {mtovar,patty}@cs.buap.mx

Resumen. En este artículo se presenta el diseño de un modelo ontológico para la búsqueda de información dentro de una institución educativa de nivel superior. Este modelo pretende brindar respuestas sobre los procesos de inscripción y liberación de las prácticas profesionales, es decir, de los requisitos que se deben llevar a cabo por el alumno para realizar este tipo de procedimientos en la institución educativa. En este trabajo se siguen las fases de diseño de la metodología propuesta por Grüninger y Fox's para la creación manual de la ontología; también se incluye la propuesta de un escenario, preguntas de competencia, definición de clases, propiedades, formalización y evaluación a través de las respuestas obtenidas por medio del lenguaje de consulta SPARQL.

Palabras clave: Metodologías, institución educativa, ontología.

Ontological Model to Represent Information about the Professional Practice in an Educational Institution

Abstract. This article presents the design of an ontological model for the search of information within a higher education institution. This model aims to provide answers about the processes of enrollment and release of professional practices, that is, the requirements that must be carried out by the student to perform this type of procedures in the educational institution. In this work the design phases of the methodology proposed by Grninger and Fox's for the manual creation of the ontology are followed; also includes the proposal of a scenario, questions of competence, the definition of classes, properties, formalization and evaluation through the answers obtained through the SPARQL query language.

Keywords. Methodologies, educational institution, ontology.

1. Introducción

Hoy en día las universidades son un sector que genera una gran cantidad de datos o información proveniente de distintas áreas de trabajo, dicha información en la mayoría de los casos no es gestionada de una forma adecuada y mucho menos es accesible para todo público que requiere de ella [6]. Hoy en día existen fuentes de información en la Web que poseen motores de búsqueda que dan respuesta a lo que solicita el usuario, sin embargo, estos motores funcionan sintácticamente, es decir, devuelven resultados que contienen elementos léxicos relacionados exactamente con los términos de la consulta, es por ello que se recurre a la Web Semántica la cual propone superar las limitaciones de la Web actual. Dados estos cambios es imprescindible unificar el contenido a través de un lenguaje común y para ello se utiliza la noción de ontología del campo de la inteligencia artificial. Gruber define una ontología como "*a formal explicit specification of a shared conceptualization*" [5]. Una ontología define una jerarquía de conceptos, relaciones, restricciones, axiomas e instancias para describir un dominio, el cual servirá para el intercambio de la información. Por otra parte el uso de las ontologías hoy en día es muy común en áreas de negocios, finanzas, internet, medicina e industria como integradoras de fuentes de datos, así como de organización y representación de conocimiento. Para este trabajo se presenta el diseño de un modelo ontológico implementado en la herramienta Protégé [7]. La finalidad es que usuarios que necesiten realizar el proceso de práctica profesional puedan consultar toda la información relacionada con dicho proceso, que consta de una serie de pasos los cuales pueden ser representados en una ontología. El documento se encuentra estructurado de la siguiente manera: En la Sección 2. se presentan algunos trabajos relacionados con la creación manual de ontológias. En la Sección 3. se presenta el planteamiento del escenario y las preguntas de competencia. En la Sección 4. se presenta el diseño de la ontología y en la Sección 5. se muestran los resultados de la evaluación. Finalmente las conclusiones se presentan en la Sección 6..

2. Trabajos relacionados

En esta sección se mencionan algunos trabajos que estan relacionados con el tema de búsqueda de información dentro de una institución educativa y creación manual de ontologías.

En Báez Bagatella, Tamborrell Hernández, Lasserre Chávez, Ramos Flores, Tovar Vidal, y Vilariño Ayala [1] se propone una ontología que contiene información relacionada con profesores, alumnos, personal administrativo, planes de estudio, infraestructura, proyectos y líneas de investigación de una institución educativa superior del estado de Puebla; se pretende que dicha información se encuentre de forma estructurada y que pueda ser consultada por humanos y otros sistemas computacionales.

El modelo ontológico propuesto se divide en planteamiento, descripción de las entidades y clases de equivalencia. En Rose Gómez, Serna Encinas y Rodríguez

Elías [9] se presenta el uso del lenguaje natural para la búsqueda de información de una organización académica, para la creación del modelo de conocimiento. Se utiliza la metodología KoFI que se compone de cuatro etapas: Identificación de fuentes de conocimiento, Identificación de tipos/temas de conocimiento, identificación de flujos de conocimiento e identificación de problemas que afecten el flujo de conocimiento. Con el análisis de las cuatro etapas se obtiene un meta modelo que es la base del modelo de conocimiento, el cual considera un repositorio de datos y un repositorio de conocimiento.

El repositorio de datos incluye una base de datos para los documentos y para el repositorio de conocimiento se usa una ontología diseñada, a través de la metodología Methontology e implementada con la herramienta Protégé. En Rosell León, Senso Ruiz y Leiva Maderos [10] se describe el desarrollo de una ontología nombrada UH-Ontology para el manejo de datos de una universidad. La metodología que se utiliza para la elaboración de UH-Ontology consta de siete pasos; 1) Determinar el dominio y el alcance de la ontología, 2) Reutilizar ontologías ya existentes, 3) Enumerar términos importantes, 4) Definir clases y jerarquía de clases, 5) Definir propiedades de clases (SLOTS), 6) Definir facetas de los slots y 7) Crear instancias.

La herramienta utilizada para la modelación e implementación de la ontología fue Protégé 4.3. En Tabares García, y Jiménez Builes [11] se presenta la construcción de una ontología para el proceso evaluativo en la educación superior, para este trabajo se utilizá la metodología de Methontology y la herramienta de Protégé 3.4.8. En Mora Arciniega y Segarra Faggioni [6] se describe la realización de un modelo ontológico para la representación de datos académicos y su publicación con tecnología semántica utilizando los conceptos de ontologías y datos enlazados, la propuesta contempla el desarrollo de una ontología para representar los datos de los planes de un curso y la publicación de estos datos mediante las mejores prácticas de Datos Enlazados. En Bravo, Martínez Reyes y Rodríguez [2] se describe un modelo ontológico para la representación de contexto académico e institucional. El sistema ofrece un contenido base que provee respuestas a las cuestiones académicas como tutorías, supervisión de tesis, además de lo relativo a ubicación de personas, bibliotecas, edificios, carreteras, horarios de clase y horarios de eventos.

El principal objetivo del modelo ontológico es mostrar las ventajas soluciones basadas en ontologías, dicho modelo esta implementado en el idioma español, el modelo ontológico integra tres ontologías las cuales son AreaGeografica, Persona y Academica. De igual manera presentan la evaluación de la ontología por medio de convertir preguntas de competencia en consultas SQRWL, dichas consultas pueden ser ingresadas en el sistema para dar respuesta a las preguntas. El modelo ontológico es implementado a través de una interfaz móvil e interfaces de escritorio donde el usuario puede interactuar con el modelo, se muestra un ejemplo donde un estudiante puede obtener la oficina y los salones de clases de un profesor. En Tovar, Flores y Reyes-Ortiz [12] se presenta un método para construir una ontología manual para la búsqueda de información sobre los trámites que debe realizar los estudiantes de una institución educativa para registrar y liberar

su servicio social. Se presentan los pasos a seguir de la metodología Grüninger y Fox's [4] y Bravo et. al., [2] para la creación manual de la ontología. En Reyes, Tovar y Vázquez [8] se presenta un método para construir una ontología manual para un programa de maestría en Ciencias de la Computación utilizando las metodologías de METHONTOLOGY, Grüninger y Fox's [4] y Bravo et. al., [2].

En este trabajo de investigación se presenta el diseño ontológico de las actividades que debe realizar un estudiante de nivel superior para registrar y liberar su práctica profesional en una institución educativa. El modelo ontológico sigue los pasos de diseño de las metodologías de Grüninger y Fox's [4] y Bravo et. al., [2].

3. Propuesta de modelo ontológico

La construcción de sistemas basados en ontologías requiere una metodología clara, puesto que no solamente se trata de la construcción de un sistema, sino también de la generación de un modelo de experimentación e investigación. Se propone la metodología de Grüninger y Fox's y Bravo et. al.[2], para la realización de este trabajo, como ya se mencionó anteriormente. La metodología de Grüninger y Fox's se desarrolla en la universidad de Toronto y está inspirada para el desarrollo de conocimiento basado en sistemas usando lógica de primer orden.

La metodología propone identificar los principales escenarios como primer paso, es decir, posibles aplicaciones en las cuales se utilizará la ontología. Como segundo paso de la metodología se propone realizar una serie de preguntas denominadas como de "competencia", las cuales serán utilizadas para determinar el alcance de la ontología, estas preguntas y sus repuestas son usadas para extraer los principales conceptos y sus propiedades, relaciones y axiomas formales de la ontología [4].

3.1. Planteamiento del escenario

El modelo ontológico brindará respuesta acerca de los trámites y requisitos que se necesitan para que un alumno de una institución educativa de nivel superior pueda realizar sus prácticas profesionales. En el caso de los requisitos el alumno podrá visualizar la lista de los documentos necesarios para cada uno de los procesos que conllevan las prácticas profesionales, los cuales son el proceso de inscripción y el proceso de liberación. De igual manera se pondrá a su alcance un listado de programas los cuales son ofertas ofrecidas por los sectores empresarial y educativo para que el alumno realice sus prácticas profesionales con ellos. En dichas ofertas se coloca todo la información relacionada como por ejemplo: nombre del programa, número de solicitantes, ubicación del lugar donde se realizarán las prácticas, nombre de la persona encargada, etc.

3.2. Preguntas de competencia

Para la realización de este modelo ontológico y siguiendo con la metodología se elaboraron una serie de preguntas denomindas "preguntas de competencia", dichas preguntas se elaboran con la finalidad de tener una idea de lo que un usuario podría realizar como búsqueda dentro del modelo ontológico, además de que nos ayudarán a identificar clases, relaciones, propiedades y axiomas para la realización de la ontología. Entre las preguntas que se elaboraron se tienen las siguientes:

1. ¿Cuál es la oficina del coordinador de práctica profesional?
2. ¿Cuántas horas son en total de práctica profesional?
3. ¿Qué documentación necesito para inscribir la práctica profesional?
4. ¿Qué documentación necesito para liberar la práctica profesional?
5. ¿Cuántos créditos tiene la práctica profesional?
6. ¿En qué periodo puedo inscribir la práctica profesional?
7. ¿Cuál es el horario de atención del coordinador de práctica profesional?
8. ¿Cuáles son los tipos de programas que existen en la práctica profesional?

3.3. Descripción de clases

En esta fase se realizó un análisis, partiendo del escenario y de las preguntas de competencia, se identificaron las clases que contendría el modelo ontológico, de igual manera se obtuvieron las relaciones y las propiedades de cada clase. En la Tabla **??** se describen las clases identificadas incluidas en el modelo ontológico y su descripción.

En las Tablas 1 y 2 se describen las propiedades de cada una de las clases, las cuales ayudaron a almacenar información para dar respuesta a las preguntas de competencia. En la Tabla 1 se muestran las propiedades de la clase "Persona", así como de sus subclases. En la tabla 2 se muestran las propiedades de las clases restantes que son "PracticaProfesional, Programa, Facultad, Tramite y Requisito", así como de sus subclases.

4. Diseño

En este apartado se presenta el diseño de la ontología propuesta, el diseño se realizó con la herramienta Protégé, la cual es un software de código abierto que permite construir una ontología mediante su interfaz que ayuda al desarrollador en el proceso [7]. En la Fig. 1 se presentan las clases con las que cuenta el modelo ontológico. Las clases tienen relaciones entre ellas, es por ello que en la Fig. 2 se muestra el diagrama de las relaciones o propiedades de objeto y se describen en la Tabla 3.

Tabla 1. Propiedades de tipo de dato de la clase Persona.

Propiedad de tipo de Dato	Dominio	Rango
tieneNombre	Persona	String
tieneDireccion	Persona	String
tieneTelefono	Persona	String
tieneSexo	Persona	String
tieneEmail	Persona	String
tieneMatricula	Alumno	String
tieneHorarioAtencion	CoordinadorPracticaProfesional	String
tieneNoTrabajador	ResponsablePrograma, CoordinadorPracticaProfesional, ProfesorResponsable	String
tieneOficina	ResponsablePrograma, CoordinadorPracticaProfesional, ProfesorResponsable	String

Tabla 2. Propiedades de tipo de dato de las clases PracticaProfesional, Programa, Facultad, Tramite y Requisito.

Propiedad de Dato	Domino	Rango
tieneNombreFacultad	Facultad	String
tieneNombreCarrera	ProgramaDeEstudio	String
tieneNombrePrograma	Programa	String
tieneNoPrograma	Programa	String
tieneFechaRegistro	Programa	String
tieneEstatus	Programa	String
tieneFechaInicio	PracticaProfesional	String
tieneFechaTermino	PracticaProfesional	String
tieneCredito	PracticaProfesional	positiveInteger
tienePeriodo	PracticaProfesional	String
tieneDuracion	Programa	String
tieneDuracionPP	PracticaProfesional	String
tienePeriodoActual	Programa	String
tieneRequisitoInscripcion	Inscripcion	String
tieneRequisitoLiberacion	Liberacion	String
tieneTipo	Programa	String

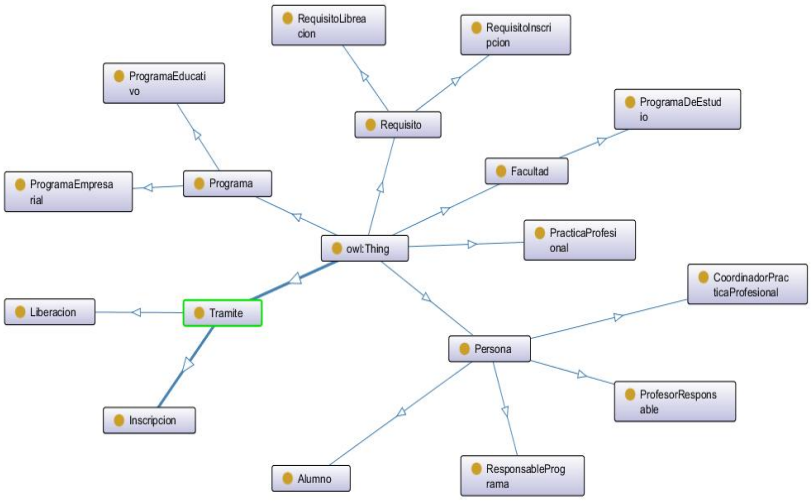

Fig. 1. Clases del modelo ontológico.

Tabla 3. Propiedades de objeto o relaciones entre clases.

Object Properties	**Domino**	**Rango**
acargoDe	ProfesorResponsable	Alumno
autoriza	CoordinadotPracticaProfesional	Liberacion, Inscripcion
dirigidoA	Programa	ProgramaDeEstudio
inscritoEn	Alumno	ProgramaDeEstudio
realiza	Alumno	Inscripcin
registraUn	ResponsablePrograma	Programa
tiene	Liberacion, Inscripcion	RequisitoLiberacion, RequisitoInscripcion
esRegistradoPor	Programa	ResponsablePrograma

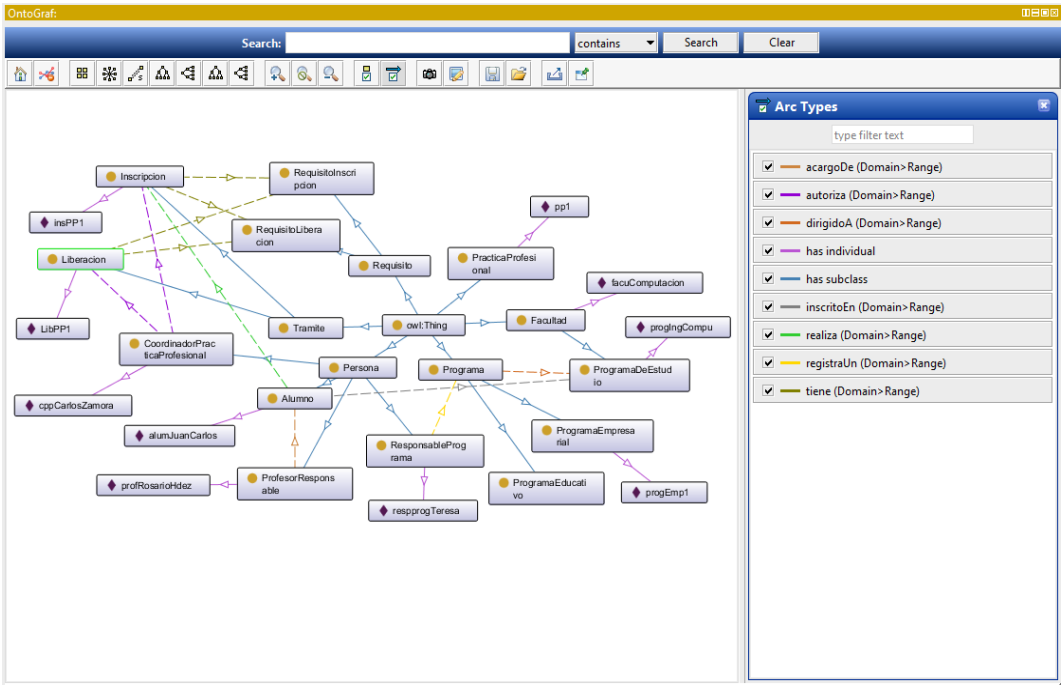

Fig. 2. Relaciones entre las clases del modelo ontológico.

5. Resultados

En esta sección se presenta la formalización de las preguntas de competencia (ver Tabla 4) las cuales se presentaron en la sección 3.2. La evaluación de la ontología se realizó a nivel de aplicación, es decir, a través de las respuestas que el sistema ontológico provee ante la formalización de las preguntas de competencia.

En este caso se utiliza el lenguaje de consulta SPARQL [3] y las respuestas de las preguntas de competencia son presentadas en las Fig. 3-10 (ver columna 3 de la Tabla 4).

6. Conclusión

La realización de esta ontología pretende que la información este organizada de forma más adecuada y que sea más fácil acceder a ella. Las ontologías son poderosas herramientas para estructurar la información, y muy utilizadas en ingeniería del conocimiento, procesamiento de lenguaje natural, inteligencia artificial, recuperación de información, análisis de sentimientos, etc. La principal contribución de este trabajo es la aplicación de una metodología para la creación manual de una ontología aplicada al dominio educativo para la búsqueda de

Tabla 4. Formalización de las preguntas de competencia y respuestas en SPARQL.

No	**Pregunta y formalización**	Consulta SPARQL
1	¿Cuál es la oficina del coordinador de práctica profesional? $\exists \$x \$y \$n \ (CoordinadorPracticaProfesional(\$x) \wedge tieneNombre(\$x, \$n)) \wedge tieneOficina(\$x, \$y))$?	Fig. 3
2	¿Cuántas horas son en total de páctica profesional? $\exists \$x \$y \ (PracticaProfesional(\$x) \wedge tieneDurationPP(\$x, \$y))$?	Fig 4
3	¿Qué documentación necesito para inscribir la práctica profesional? $\exists \$x \$y \ (Inscripcion(\$x) \wedge tieneRequisitoInscripcion(\$x, \$y))$?	Fig. 5
4	¿Qué documentación necesito para liberar la práctica profesional? $\exists \$x \$y \ (Liberacion(\$x) \wedge tieneRequisitoLiberacion(\$x, \$y))$?	Fig. 6
5	¿Cuántos creditos tienen la práctica profesional? $\exists \$x \$y \ (PracticaProfesional(\$x) \wedge tieneCredito(\$x, \$y))$?	Fig. 7
6	¿En qué periodos puedo inscribir la práctica profesional? $\exists \$x \$y \ (PracticaProfesional(\$x) \wedge tienePeriodo(\$x, \$y))$?	Fig. 8
7	¿Cuál es horario de atención del coordinador de la práctica profesional? $\exists \$x \$y \$n \ (CoordinadorPracticaProfesional(\$x) \wedge tieneNombre(\$x, \$n)) \wedge tieneHorarioAtencion(\$x, \$y))$?	Fig. 9
8	¿Cuáles son los tipos de programas que existen en la práctica profesional? $\exists \$x \ (Programa(\$x))$?	Fig. 10

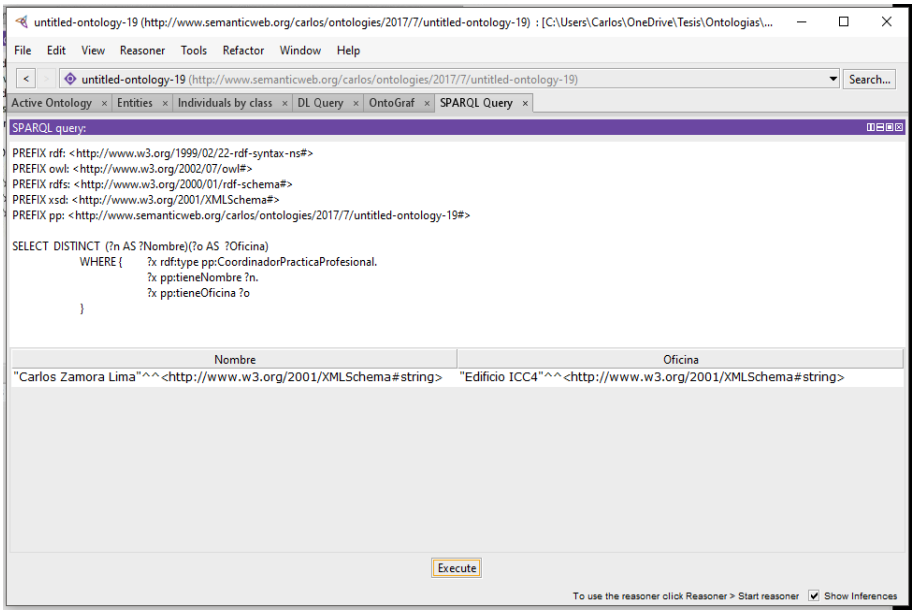

Fig. 3. Respuesta a la pregunta 1.

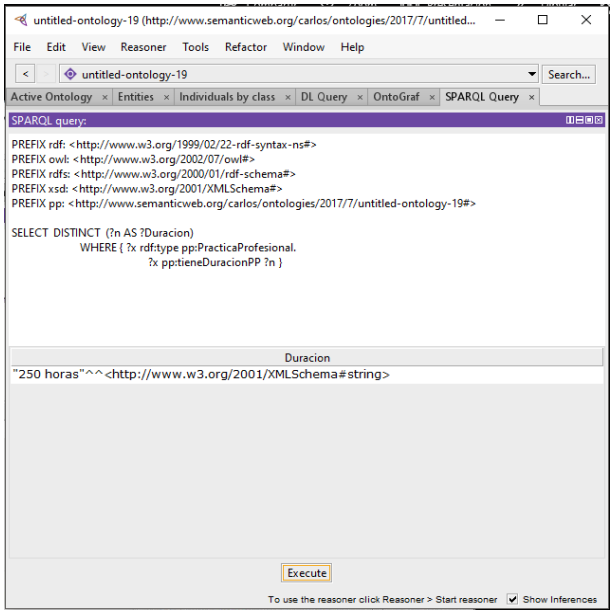

Fig. 4. Respuesta a la pregunta 2.

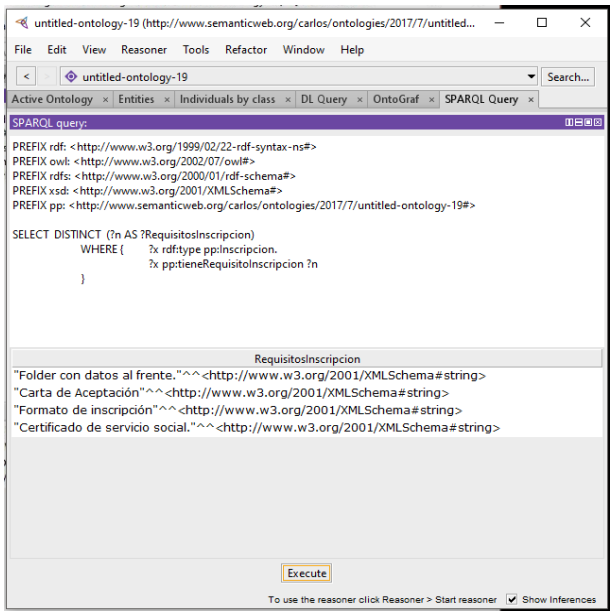

Fig. 5. Respuesta a la pregunta 3.

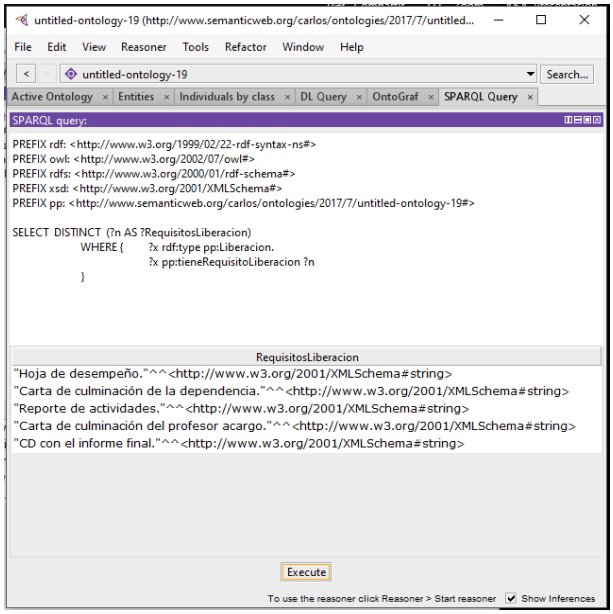

Fig. 6. Respuesta a la pregunta 4.

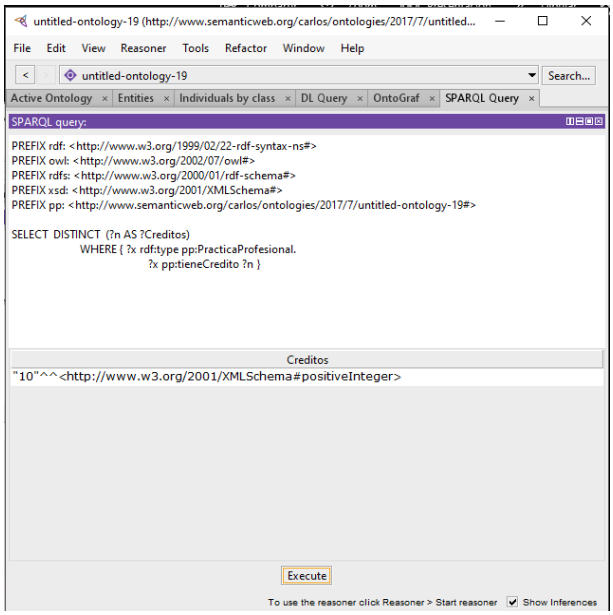

Fig. 7. Respuesta a la pregunta 5.

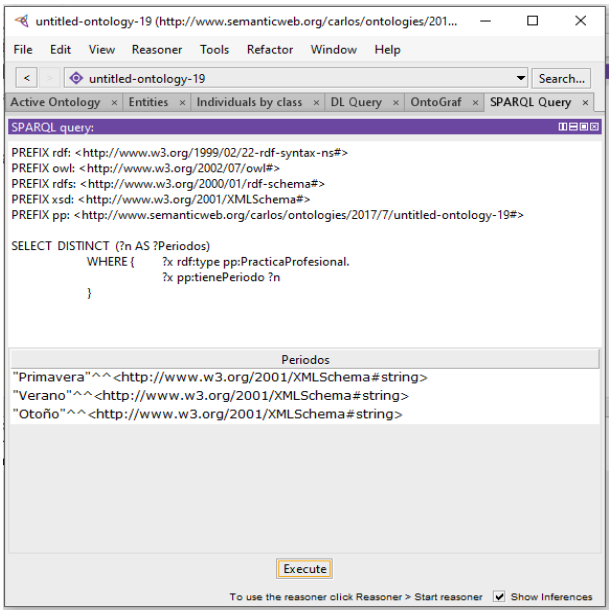

Fig. 8. Respuesta a la pregunta 6.

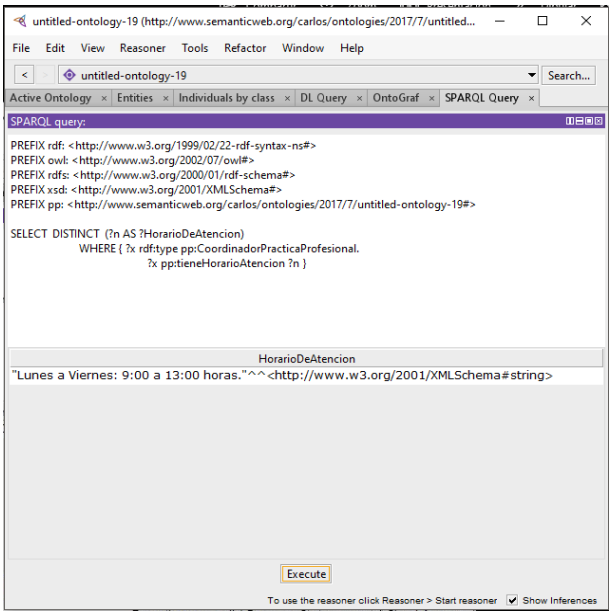

Fig. 9. Respuesta a la pregunta 7.

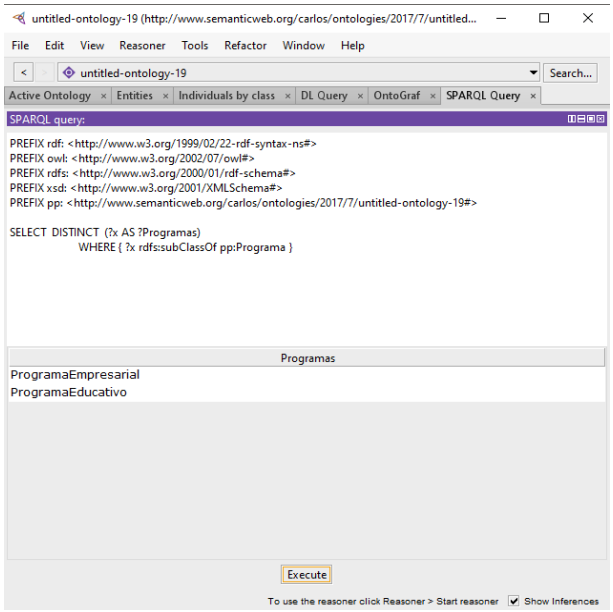

Fig. 10. Respuesta a la pregunta 8.

información acerca de trámites o servicios como es la práctica profesional a nivel superior. Como trabajo a futuro se pretende la creación de una aplicación que utilice la ontología para hacer el poblado autómatico a nivel de instancias de cada clase definida en la ontología y la presentación de las respuestas de las preguntas de competencia al usuario final.

Agradecimientos. Esta investigación es apoyada por el Fondo Sectorial de Investigación para la Educación, proyecto Conacyt CB/257357. Por el proyecto ID 00478 VIEP-BUAP y por el proyecto PRODEP-SEP ID 00570 (EXB-792) DSA/103.5/15/10854.

Referencias

1. Baez Bagatella, J.A., Tamborell Hérnandez, A., Lasserre Chávez, H., Ramos Flores, O., Tovar Vidal, M., Vilariño Ayala, D.: Un modelo ontológico para representar la organización de una unidad educativa. In: Avances recientes en ciencias computacionales - CICOMP 2016 (2016)
2. Bravo, M., Martínez Reyes, F., Rodríguez, J.: Representation of an academic and institutional context using ontologies. Reseach in Computing Science 87, 9–17 (2014)
3. DuCharme, B.: Learning SPARQL querying and updating with SPARQL 1.1. O'Reilly Media, Sebastopol, CA (2013)

4. Gómez Pérez, A., Fernández López, M., Corcho, O.: Ontological Engineering with examples from the areas of knowledge Management e-Commerce and the Semantic Web. Springer (2004)
5. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition 5(2), 199–220 (Jun 1993), http://dx.doi.org/10.1006/knac.1993.1008
6. Mora Arciniega, M.B., Segarra Faggioni, V.: Modelo ontológico para la representación de datos académicos y su publicación con tecnología semántica. Opción 10, 267–282 (2016)
7. Musen, M.A.: The protégé project: A look back and a look forward. AI Matters 1(4), 4–12 (Jun 2015), http://doi.acm.org/10.1145/2757001.2757003
8. Reyes, C., Tovar, M., Vázquez, S.: Ontology for the description of a masters degree program in computer sciences. In: Proceedings of the XVIII International Conference on Human Computer Interaction. pp. 12:1–12:4. Interacción '17, ACM, New York, NY, USA (2017), http://doi.acm.org/10.1145/3123818.3123836
9. Rose Gómez, C.E., Serna Encias, M.T., Rodríguez Elías, O.M.: Uso del lenguaje natural para la búsqueda de respuestas en un sistema de gestión de conocimiento académico. Tecnologías emergentes y avances de la computación en México Primera Edición, 155–160 (2016)
10. Rosell León, Y., Senso Ruiz, J.A., Leiva Mederos, A.A.: Diseño de una ontología para la gestión de datos heterogéneos en universidades. Información en Ciencias de la Salud 27, 545–567 (2016)
11. Tabares García, J.J., Jiménez Builes, J.A.: Ontología para el proceso evaluativo en la educación superior. Revista Virtual Universidad Cat'olica del Norte 42, 68–79 (2014)
12. Tovar, M., Flores, J.C., Reyes, J.: An ontology for representing information over social service in an educational institution. In: Conference Special Session on Knowledge Discovery and Cloud Computing Applications. pp. 391–399 (01 2017)

ISSN 1870-4069

A Semantic Proposal for Semiautomatic Corpus Creation in the Pedagogic Domain

Yuridiana Aleman, María Somodevilla, Darnes Vilariño

Benemérita Universidad Autónoma de Puebla,
Faculty of Computer Science, Puebla, Mexico

{yuridiana.aleman,mariajsomodevilla,dvilarinoayala}@gmail.com

Abstract. In this research is introduced a methodology for the ontology automatic construction for the pedagogical domain based on the definition of two lexical resources: a domain corpus and, the use of a domain dictionary. In this research, the pedagogical corpus was building manually, where aspects such as learning strategies, intelligences types and learning styles were included. After that, the corpus is processed for extracting a list of concepts linked to the domain predefined classes. Each concept is searched in five dictionaries, including synonyms. This process is automatically performed, removing close words, words with low frequency. As result of this process, is building a domain dictionary which include concepts related with each domain class. The preliminary results show an affinity between the corpus and the dictionary, which is a very important resource for an ontology definition process. As future work, these resources will be used for detect classes in a ontology learning process for the pedagogical domain using machine learning techniques.

Keywords: Dictionary, synonyms, pedagogy domain, class detection, learning styles, intelligences types, learning strategies, corpus.

1 Introduction

In recent years, the available information has increased exponentially and the classic techniques of information retrieval cannot resolve problems such as semantic questions and language interpretation. Ontologies can be used for purposes such as structure knowledge in taxonomies, vocabulary manage, natural language processing applications, searches, recommendation systems, and e-learning among others [5]. The ontology learning process integrates the class detection, creation, populate and evaluation of ontologies. This process is applied in several domains focusing on one or more steps such has education, tourism, financial, among others ([9,6,7,16]).

Ontology is a formal, explicit specification of a shared conceptualization. Their classes, relationships, constraints and axioms define a common vocabulary to share knowledge [11]. The ontology learning process needs resources with specific characteristics and compatible with the domain to research. This paper is focused on the previous step of ontology learning: resources construction for class

detection in the future. The resources are a dictionary with principal concepts and a corpus with papers published, the research will be realized for pedagogical domain in Spanish language. Pedagogical domain is extensive, thus the research is focused in the creation of support tools for the teachers in a class. Three topics was research: learning styles, intelligences types and learning strategies in the class.

The article is organized in 7 sections following described. Section 2 introduces the problem as well as the relevance to the selected domain. Section 3 presents the proposed methodology for the resources construction, and the sections 4 and 5 describes the resources. Preliminary results of quantitative analysis is shown in section 6. Finally, section 7 presents conclusions and future work for the research.

2 Related Work

In this section, the works about class detection and corpus created are analyzed.

Ontologies for the Use of digital learning Resources and semantic Annotations on Line (OURAL) project is presented in [9], the project includes people from several disciplines (educational science, computer science, and cognitive psychology) building e-learning services. The authors present the extracted class using Natural Language Process techniques in unstructured texts about learning situation. Educational domain was also analyzed in [6], but its application was into Chinese language. The authors analyzed three features of Chinese language: Coupling, Domain Relevancy and Domain Consensus; these features be modeled and integrated to evaluate the terms. Snow ball technique was used to acquire hyponymy relation, and HowNet-based method for extract general relation.

Others works like [17] present methods for semi-automatic class extraction using a database of Spanish verbs, diathesis alternations and syntactic-semantic schemes (ADESSE tool) [7], where the semantic extracted patterns are the classes. This methodology was applied in educational domain and replicated in financial domain in [16]; in both works, the class extraction was completed with the domain expert opinion. A method for class extraction using linguistic patterns and NLP metrics such as morphological labeling is presented in a recent research [15].

The researches which report ontologies manually constructed is based in the domain and evaluation, since the domain used for this proposal, mainly analyzes the works in pedagogical domain. In [22] an ontology for interaction between students and teachers for English language teaching is introduced. On the other hand, [13] proposes an ontology for the internet learning process. In both works is defined an ontology for each entity in the learning process, and the evaluation is conducted with a manual process supervised for domain experts. Other researchers are focused on online education ([21,3,4] and recently [12]) where ontologies are manually defined from XML resources available in the Internet, and the evaluation is a manual process too. An ontology created from CASE diagrams for on-line education is presented in [2]; its evaluation is

addressed by experts in a manual process. In this researches, the focus is the construction step where the class are extracted manually.

There are works such as [20] focused on automatic learning; in this paper, an ontology based on the Internet of Things used in a classroom is created, considering the student intelligences. The ontology creation process from the courses information offered in advanced levels is explained in [1], where students can choose courses according with their academic background. Both works present the structure, information, and hierarchy of the classes in a manual way.

Some researches are focused in the corpus creation, but in different domains. [10] is focused on the creation of the relevant linguistic corpus written in Serbian language. The focus is the sentiment analysis of student generated contents on higher education. In [19] the problem of creating a reference corpus for the classification of news items in fine grained multi-label scenarios was analyzed. The authors propose a semiautomatic approach for creating a reference corpus that uses three auxiliary classification methods: Support Vector Machines, Nearest Neighbor Classifiers and another based on a dictionary.

3 Methodology

Figure 1 shows the steps proposed, where two parallel methods was worked: corpus and dictionary. For the corpus construction some academic researches was used with two principal characteristics: focused in social sciences (pedagogy) and contain papers written in Spanish language. Secondly, papers about the principal class was extracted and joined in a initial corpus.

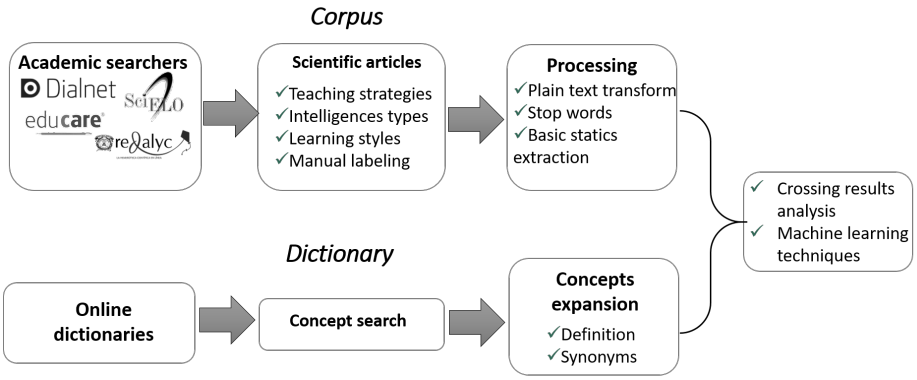

Fig. 1. Corpus built process.

In first step, four academic researches was choose:

- **Dialnet**[1]**:** Is is a project started in Rioja University. The site web collects and gives access to documents published in Spain in any language about Hispanic topics.
- **Scientific Electronic Library Online (SciElo)**[2]**:** It is a model for cooperative electronic publication of scientific journals in Internet. Its goal is to respond the scientific communication needs in developing countries (Latin America and Caribbean)
- **Educare Electronic Journal**[3]**:** Is a quarterly international electronic magazine of Costa Rica University. The journal diffuse the science production and develop the academic analysis in the education domain.
- **Redalyc**[4]**:** It is a online information system for scientific magazines in Latin America, the Caribbean, Spain and Portugal.

The dictionary was constructed using 5 principal resources: 2 pedagogical dictionaries, definition of Spanish Real Academy, encyclopedic dictionary (print edition) and online synonyms dictionary. The complete process are explained in section 5.

4 Corpus Description

Corpus represents a set of data or a collection of texts used for linguistic analysis and Natural Language Processing. It is composed of words that are not designed by linguists, but the authentic words that occur naturally in written or spoken language. From the aspect of corpus linguistics, corpus must meet two basic conditions: representative and balanced [18]. For this reason, the corpus is created using a reality sample; in this case, with published papers in the domain. Three principals topics was used for papers research:

1. **Learning styles**: Project the way of which a person learn. However, there exist alternatives about how is possible to learn concepts and processing information by humans. In different works, have been proposed several theories for describe the different types of learning. This work adopt as reference the Neuro-Linguistic Programming model (NLP), which implement a technique that allow to improve the communication level between instructors and students. Three styles were analyzed: visual, auditory and kinesthetic.
2. **Intelligences types**: An intelligence is the ability to solve problems, or to create products, that are valued within one or more cultural settings [8]. Humans have a capacities range and potentials that can be employed in productive ways (together or separately). This idea origined the multiple intelligences theory. The types of intelligence identified in [8]; linguistic, logical-mathematical, musical, bodily, spatial, interpersonal and intrapersonal.

[1] https://dialnet.unirioja.es/
[2] http://www.scielo.org/php/index.php?lang=es
[3] http://www.revistas.una.ac.cr/index.php/EDUCARE/index
[4] http://www.redalyc.org/home.oa

3. **Teaching strategies**: From teaching-learning methodologies described in literature, the Model-Eliciting Activities (MEA) will be analyzed, starting with the theory of mental processing [14]. These methodologies have been used from elementary education to professional education, with the propose to identify the way in which students learn concepts, and establish units, indicators and tools for conduct the investigation analysis.

Table A (appendix section) shows the papers selected for the corpus, including the publication year, first author and the magazine. Some thesis was used, but only the theoretical framework. The papers text was extracted and preprocessed in plain text. The result of this process was a corpus A with 21 instances, where each instance is a paper. A can be described such as $A = \{K, T, C\}$ where:

- K is a paper key and a numeric attribute $\{1...21\}$.
- T is the all paper text, including the title and abstract. in this texts, stops words, numbers and words with length less to 2 letters was deleted.
- C is the instance class, this is a nominal attribute according to the principal topic in the paper. $C = \{LearningStyle, IntelligenceType, LearningStrategy\}$. Each paper was manually labeled according its title; the corpus was balanced, thus exists 7 instances for each class.

Table 1 shows the vocabulary frequency in each class, this analysis was realized afterward initial preprocessing.

Table 1. Corpus Vocabulary.

Class	Words	Vocabulary
LearningStyle	13,228	3,528
LearningStrategy	19,264	4,450
IntelligenceType	24,609	6,162
Total	57,081	9,625

IntelligenceType class has more vocabulary, but this class contain 7 principal concepts, the, the difference between the class is justified. Analyzed the total; the class shared many words.

5 Dictionary Description

The construction process was similar to corpus construction. First a initial list of class and components were created manually, this list contains the subdivision of the principal class. The words were the follows:

actividad	instrumentos	matemático	prototipo
auditivo	inteligencia	modelo	realidad
autoevaluación	interpersonal	mental	reutilización
corporal	intrapersonal	musical	transmisión
espacial	kinestésico	naturalista	visual
aprendizaje	lingüística		

Secondly, the concepts was searched in the resources mentioned in section 3. Then, a initial dictionary D_i of 22 concepts was created. D_i can be described such as $D_i = \{W, C_1, C_2, C_3, C_4, S\}$, where:

- W is the principal concept (initial list).
- C_1, C_2, C_3, C_4 are the definition in each of the dictionaries used. Some words, do not have definition in some resource. The texts was preprocessed deleting stops words, punctuation marks and words with length less to 3 letters.
- S a list of word synonyms. The list was preprocessed deleting the punctuation marks in plain text.

D_i was expanded using the words of definition and synonyms list for obtain a dictionary D_f. This dictionary is a simple list than only contain the concepts. Figure 2 shows an example with a generic word (no included in D_i).

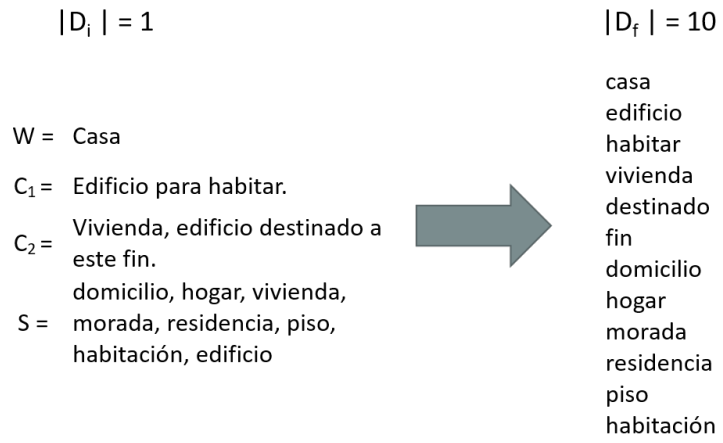

Fig. 2. Example of dictionary expansion.

Applied the expansion process, the list increased to 1,047 words. A final process was applied and the words with unitary frequency was deleted; then $\|D_f\| = 337$.

6 Preliminary Results

In this section, the result of crossing process was analyzed. Each of the D_f concepts was searched in A and the coincidences were counted. The number of matches are presented as follows:

LearningStyle	188
LearningStrategy	219
InteligenceType	239
Total Vocabulary	258

The IntelligenceType class has more words, but it is for the vocabulary size. In total, the number of coincidences is 258. The words of D_f that no appear in the corpus was deleted, thus $\|D_f\| = 258$. Finally, Table 2 shows the words of A with more frequency that appear in D_f

Table 2. Words in corpus with more frequency.

Word	Frequency
aprendizaje	895
estilos	420
inteligencia	397
inteligencias	272
educacion	256
actividades	243
proceso	227
ser	219
manera	210
asi	210
ensenanza	207

The words presented are related with the definition of the principal class, thus, is possible start with the process for the ontological learning. Also, the corpus was analyzed using supervised classification algorithms. Table 3 shows the accuracy for random forest and support vector machine algorithms using the vocabulary and the dictionary words like attributes. The results are better with the dictionary words with 81% of accuracy, thus, using the random forest algorithm, the corpus and the dictionary are stable with the main classes for applied the ontology learning process in the future.

7 Conclusions and Future Work

In this paper the construction process of lexical resources for pedagogical domain was presented. The quantitative analysis was added, and the preliminary results shows affinity between the resources (corpus and dictionary).

Table 3. Words in corpus with more frequency.

Algorithm	Vocabulary	Dictionary
Random forest	0.619	0.809
SMO	0.666	0.761

In future word, an analysis using a tagger in Spanish will be realized for to integrate words with similar sense. Then, it is necessary add to the dictionary hyperonyms for complete the semantic analysis. This research is the first step for applied the ontological learning in pedagogical domain, thus, the future word involve the ontology construction process using a semi supervised approach.

Acknowledgements. We would like to thank Vicerrectoria of Research and Graduate Studies, Benemérita Universidad Autónoma de Puebla for supporting this work through the project Model of Teaching-Learning Process applying Ontological Engineering.

References

1. Ameen, A., Khan, K.U.R., Rani, B.P.: Creation of ontology in education domain. In: 2012 IEEE Fourth International Conference on Technology for Education. pp. 237–238 (July 2012)
2. Bagiampou, M., Kameas, A.: A use case diagrams ontology that can be used as common reference for software engineering education. In: 2012 6th IEEE International Conference Intelligent Systems. pp. 035–040 (Sept 2012)
3. Dai, X., Li, X.: Study of learning source ontology modeling in remote education. In: 2010 International Conference on Multimedia Technology. pp. 1–4 (Oct 2010)
4. Du, L., Zheng, G., You, B., Bai, L., Zhang, X.: Research of online education ontology model. In: 2012 Fourth International Conference on Computational and Information Sciences. pp. 780–783 (Aug 2012)
5. El-Ansari, A., Beni-Hssane, A., Saadi, M.: A multiple ontologies based system for answering natural language questions. In: Rocha, Á., Serrhini, M., Felgueiras, C. (eds.) Europe and MENA Cooperation Advances in Information and Communication Technologies. pp. 177–186. Springer International Publishing, Cham (2017)
6. Fu, J., Jia, K., Xu, J.: Domain ontology learning for question answering system in network education. In: 2008 The 9th International Conference for Young Computer Scientists. pp. 2647–2652 (Nov 2008)
7. García-Miguel, J.M., Vaamonde, G., Domínguez, F.G.: Adesse, a database with syntactic and semantic annotation of a corpus of spanish. In: Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (may 2010)
8. Gardner, H.: Estructuras de la Mente. Fondo de Cultura Económica (Sep 2001), `http://educreate.iacat.com/Maestros/Howard_Gardner_-_Estructuras_de_la_mente.pdf`

9. Grandbastien, M., Azouaou, F., Desmoulins, C., Faerber, R., Leclet, D., Quenu-Joiron, C.: Sharing an ontology in education: Lessons learnt from the oural project. In: Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007). pp. 694–698 (July 2007)
10. Grljević, O., Bošnjak, Z.: Development of serbian higher education corpus. In: Computational Intelligence and Informatics (CINTI), 2015 16th IEEE International Symposium on. pp. 177–181. IEEE (2015)
11. Guarino, N., Masolo, C., Vetere, G.: Ontoseek: content-based access to the web. IEEE Intelligent Systems and their Applications 14(3), 70–80 (May 1999)
12. Hssina, B., Bouikhalene, B., Merbouha, A.: An ontology to assess the performances of learners in an e-learning platform based on semantic web technology: Moodle case study. In: Europe and MENA Cooperation Advances in Information and Communication Technologies, pp. 103–112. Springer (2017)
13. Hu, J., Li, Z., Xu, B.: An approach of ontology based knowledge base construction for chinese k12 education. In: 2016 First International Conference on Multimedia and Image Processing (ICMIP). pp. 83–88 (June 2016)
14. Johnson-Laird, P.N.: Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Harvard University Press, Cambridge, MA, USA (1983)
15. Kang, Y.B., Haghighi, P.D., Burstein, F.: Cfinder: An intelligent key concept finder from text for ontology development. Expert Systems with Applications 41(9), 4494–4504 (2014), `http://www.sciencedirect.com/science/article/pii/S0957417414000189`
16. Ochoa, J.L., Hernández-Alcaraz, M.L., Almela, A., Valencia-García, R.: Learning semantic relations from spanish natural language documents in the financial domain. In: Proceedings of the 3rd International Conference on Computer Modeling and Simulation, held at Mumbai, India. Chengdu: Institute of Electrical and Electronics Engineers, Inc. pp. 104–108 (2011)
17. Ochoa Hernández, J.L.: Desarrollo de una metodología para la construcción automática de ontologías en español a partir de texto libre. Ph.D. thesis, Departamento de Ingeniería de la información y las comunicaciones. Universidad de Murcia (2011)
18. Pustejovsky, J., Stubbs, A.: Natural Language Annotation for Machine Learning. A Guide to Corpus Building for Applications. O Reilly Media, USA (2013)
19. Teixeira, J., Sarmento, L., Oliveira, E.: Semi-automatic creation of a reference news corpus for fine-grained multi-label scenarios. In: Information Systems and Technologies (CISTI), 2011 6th Iberian Conference on. pp. 1–7. IEEE (2011)
20. Uskov, V., Pandey, A., Bakken, J.P., Margapuri, V.S.: Smart engineering education: The ontology of internet-of-things applications. In: 2016 IEEE Global Engineering Education Conference (EDUCON). pp. 476–481 (April 2016)
21. Wu, H.: Research of internet education system based on ontology. In: 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery. vol. 4, pp. 602–605 (Oct 2008)
22. Zhu, F., Fok, A.W.P., Ip, H.H.S., Cao, J.: Engonto: Integrated multiple english learning ontology for personalized education. In: 2008 International Conference on Computer Science and Software Engineering. vol. 5, pp. 210–213 (Dec 2008)

A Corpus description

Table 4. Papers list of initial corpus.

Author(Year)	Magazine Congress	Title
Morales, R. (2017)	Campus Virtuales	Inclusión de estilos de aprendizaje como estrategia didáctica aplicada en un AVA
Gómez, E. (2017)	Revista virtual Universidad Católica de Chile	Estilos de aprendizaje en universitarios, modalidad educación a distancia
Salas, J. (2014)	Revista Electrónica Educare	Estilos de aprendizaje en estudiantes de la Escuela de ciencias del Movimiento Humano y Calidad de Vida, Universidad Nacional Costa Rica
Bañuelos, A. (2014)	Esseñanza e Investigación en Psicología	Estilos de aprendizaje y su transformación a los largo de la trayectoria escolar
Cala, R. (2014)	Estilos de aprendizaje	Determinación de los estilos de aprendizaje de estudiantes de 1er curso de ing. Industrial y electrónica de la universidad técnica del Norte, Ibarra Ecuador
Tocci, A (2013)	Estilos de aprendizaje	Estilos de aprendizaje de los alumnos de ingeniería según la programación neuro linguïstica
Aragón, M. (2009)	CPU-e Revista de investigación Educativa	Diagnóstico de los estilos de aprendizaje en los estudiantes: Estrategia para elevar la calidad educativa
Álvarez, S. (2015)	TECTZAPIC Revista de divulgación científica y tecnológica	Actividades reveladoras del pensamiento que activan la reflexión matemática en un grupo de cálculo diferencial
Sánchez, G. (2015)	Proceedings of the Satellite conference of the International Association for Statistical Education	Implementación y transadaptación de actividades reveladoras del pensamiento (MEAS) en la enseñanza de estadística en el niuvel medio superior y a nivel superior. Un caso de estudio
Hernandez, F. (2013)	VII Congreso de Investigación, Innovación y Gestión Educativas	Actividades reveladoras del pensamiento: Una alternativa para desarrollar competencias matemáticas
Osorio, G. (2012)	Biblioteca Online ITESM (Tesis)	Las actividades reveladoras del pensamiento como estrategia de evaluación formativa en el área de matemáticas en el quinto grado de educación primaria
Alvarez, S. (2012)	Biblioteca Online ITESM (Tesis)	Actividades reveladores del pensamiento que activan el pensamiento matemático de los estudiantes dentro del proceso de las aplicaciones de la derivada que utilizan máximos y mínimos
Valdespino, E. (2011)	Biblioteca Online ITESM (Tesis)	Análisis de las respuesta de los estudiantes al trabajar una MEA con tablas de multiplicar en educación básica
Domínguez, A. (2009)	X Congreso Nacional de Investigación Educativa	Actividades reveladoras del pensamiento: más que una forma de aprendizaje activo
Garzón, A. (2016)	Revista de pedagogía	La i ntegración TIC-Inteligencias múltiples (IM): Una oportunidad de cambio en el proceso educativo
Barraza, R. (2016)	Actualidades investigativas en Educación	Rendimiento académico y autopercepción de inteligencias múltiples e inteligencia emocional en universitarios de primera generación
Mainieri, A. (2015)	Actualidades investigativas en Educación	Conocimientos teóricos y estrategias mtodológicas que emplean docentes de primer ciclo en la estimulación de las inteligencias múltiples
Inciarte, N. (2012)	Multiciencias	Inteligencias múltiples en la formación de investigadores
Juárez, J. (2010)	Investigación y Posgrado	Inteligencias Múltiples: Una innovación pedagógica para potenciar el proceso enseñanza aprendizqje
Paniagua, L. (2008)	Revista Electrónica Educare	La teoría de las inteligencias múltiples en la práctica docente en educación preescolar
Guzman, B. (2005)	Revista de investigación	Las inteligencias múltiples en el aula de clases

ISSN 1870-4069

Analysis of EEG Signal Processing Techniques based on Spectrograms

Ricardo Ramos-Aguilar, J. Arturo Olvera-López, Ivan Olmos-Pineda

Benemérita Universidad Autónoma de Puebla,
Faculty of Computer Science, Puebla, Mexico

ricramosa1@gmail.com,{aolvera, iolmos}@cs.buap.mx

Abstract. Current approaches for the processing and analysis of EEG signals consist mainly of three phases: preprocessing, feature extraction, and classification. The analysis of EEG signals is through different domains: time, frequency, or time-frequency; the former is the most common, while the latter shows competitive results, implementing different techniques with several advantages in analysis. This paper aims to present a general description of works and methodologies of EEG signal analysis in time-frequency, using Short Time Fourier Transform (STFT) as a representation or a spectrogram to analyze EEG signals.

Keywords. EEG signals, spectrogram, short time Fourier transform.

1 Introduction

The human brain is one of the most complex organs in the human body. It is considered the center of the human nervous system and controls different organs and functions, such as the pumping of the heart, the secretion of glands, breathing, and internal temperature. Cells called neurons are the basic units of the brain, which send electrical signals to control the human body and can be measured using Electroencephalography (EEG).

Electroencephalography measures the electrical activity of the brain by recording via electrodes placed either on the scalp or on the cortex. These time-varying records produce potential differences because of the electrical cerebral activity. The signal generated by this electrical activity is a complex random signal and is non-stationary [1]. Signals are normally composed of four brainwave groups: Delta, Theta, Alpha, and Beta [2]. Each band has different frequencies and amplitudes related to activities in the human body.

Currently, the analysis of EEG signals is very important due to the information obtained from the signal, which can help physicians to recognize brain issues such as epilepsy, Alzheimer's, seizure disorder, attention deficit disorder, anxiety disorder, fetal alcohol syndrome, and autism, among others [3,4,5,6,7]. The collected data also allows the brain to interact with machines or devices without physical contact, a process commonly called Brain Computer-Interfaces (BCI), which is the basis for different

applications, e.g., assistive technologies for those with severe motor disabilities [8,9]. Finally, EEG signals are proposed as an alternative to biometric applications, as they are unique and cannot be forged [7].

There are different approaches to analyzing EEG with two of the most common being time and frequency domain. However, this analysis is not unique since it depends of the application; besides, considering that non-stationary signals are considered and different artifacts are added, classifying mental stages is not a trivial problem with opportunity to do research in this area. Another problem with a big challenge is dimensionality, since the information obtained for analysis, does not usually use only a few channels of acquisition[10]. Nevertheless, not all the information obtained is useful; it could present redundant data and increase computational time [3]. Feature extraction and selection can reduce the data dimension, in addition to showing relevant patterns of EEG signals associated with the brain activity, which may reflect good performance in classification [11,12], showing other opportunity area to research.

This paper analyzes various techniques applied for processing EEG signals, showing their characteristics mainly spectrograms-based methods. The outline of this paper is as follows: The phases of EEG analysis are discussed in the following section. An overview of reviewed works is presented in the Discussion section. Finally, conclusions are stated in the last section.

2 EEG Signal Analysis Based on Spectrograms

EEG signal analysis is commonly based on three modules or phases: Artifact removal or preprocessing, Feature Extraction, and Classification, as shown in Figure 1. The first phase is preprocessing, which cleans the signal of artifacts stored during acquisition; the next phase, feature extraction, retrieves the relevant features from a previously obtained spectrogram when SSTF is applied to the clean signal; finally, in the last phase, features are provided to classifiers to construct a model for the analysis of future cases.

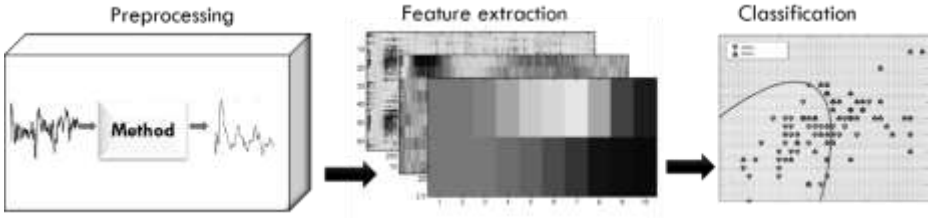

Fig. 1. Phases of EEG Signal Analysis.

2.1 Preprocessing

The preprocessing phase is the first step, which has two objectives: remove artifacts and filter data such as blinking, heartbeats, and other effects [4,13]. These can be implemented in the acquisition (phase) through hardware or by software-based techniques. Some well-known techniques for preprocessing EEG signals are Common Spa-

tial Patterns (CSP), Principle Component Analysis (PCA), Common Average Referencing (CAR), Surface Laplacian(SL), Independent Component Analysis(ICA), Adaptive Filtering, Digital Filter, and others [4,14].

The approaches based on CSP construct spatial filters that both maximize the variance of one task and minimize the variance of another [4]. CSP is a discriminative user-specific filtering technique for enhancing the signal, which detects patterns within the EEG signals by incorporating the spatial information of the EEG. It uses covariance matrices as the basis and seek a discriminative subspace, such that the variance for one class is maximized and the variance of the other is minimized simultaneously [14]. PCA converts a number of correlated variables into a set of values of uncorrelated variables called principal components.

The primary axis or the first principal component is calculated such that it accounts for the largest amount of variability in the data. Subsequent components are calculated and account for the direction of the remaining variability, but in decreasing order of the amount of variability in the data, representing the direction of the next largest variation. As the transformed data have most of the variation in the first components, the remaining components can be ignored to decrease the dimensionality [14]; CAR removes the noise by subtracting the common activity from the position of interest to improve the Signal-to-Noise Ratio (SNR), which measures the EEG signal strength relative to background noise. SL is a method with the same objective as CAR: to improve SNR and allow it to view the EEG signal with a high spatial resolution [4,14]. The CAR method re-references the signal to a common average across all the neurosensors by subtracting the mean value of the signal in all electrodes from each sample. This mean represents an estimate of the activity at the reference site, and subtracting this average produces a de-referenced solution.

Contrary to CAR, the Surface Laplacian method is derived at a specific electrode position by subtracting a combination of the signals from a set of electrodes surrounding the central electrode signal [5]; ICA is a statistical and computational method for revealing the hidden sources/components that underlie sets of random variables, such as EEG signals. ICA assumes that the unknown underlying sources are independent of each other and have been linearly combined to form a mixed signal. This returns the independent components when this independence assumption is correct [5]. Adaptive filtering can modify signal properties according to the specific characteristics of the EEG, an efficient method for solving the problem of signals and interferences with overlapping spectra [14]. Digital filters are commonly used in artifact processing, where filters are implemented as low-pass, high-pass, band-pass, and band stop; these need to select the appropriate frequency to filter noise and artifacts [4]. Table 1 shows different characteristics of the methods described in [4,14]; most of these methods focus on temporal (CSP, CAR, PCA, SL, ICA) and spatial (Fourier Analysis, Autoregressive) filters. One important difference between filters in the response during online analysis is the time, as spatial filtering is better than temporal filtering with regard to time [4]. These different methods depend on the application, data dimension, and technology used in acquisition.

Table 1. Characteristics of preprocessing methods.

Method	Characteristics
CAR	Improves SNR and outperforms all reference methods; incomplete coverage causes mistakes.
ICA	Computationally efficient and high performing for large data sizes; cannot be applied to some cases and requires more calculations for decomposition.
SL	Robust against artifacts generated at regions uncovered by electrodes, sensitive to artifacts and spline patterns.
PCA	Reduces dimension and works better than ICA.
CSP	Processes correctly motor imagery data and needs multiple electrodes (>64)
Adaptive Filter	Works well for signals with overlapping spectra and needs two signals and one reference signal.
Digital Filter	Easily removes noise, requires multiple frequencies.

2.2 Feature Extraction and Classification

The second phase in EEG signal analysis is feature extraction, where features of the signal are obtained using different signal processing techniques, such as Fast Fourier Transform (FFT), Principal Component Analysis (PCA), Wavelet Transformations (WT), Auto Regressive (AR), and others [4]. Analysis in time or frequency domain offers only time/frequency and amplitude information. The aforementioned techniques are commonly used in both domains. The Time-Frequency domain allows extracting information in the two domains simultaneously; EEG analysis is based on the time-frequency image processing technique or spectrogram, a technique commonly used in Short Time Fourier Transform, which maps the signal into a two-dimensional function of frequency and time [2]. This section shows a review of the literature on extracting features of EEG signals using STFT.

EEG signals in time-frequency domain are retrieved using the spectrogram, by applying a Short Time Fourier Transform to the signal. STFT is applied to partition the EEG signal into several segments of short-time signals by shifting the time window with some overlapping [15], a process called windowing. Depending on the time windowing function $w[n]$, a spectrogram is classified as a narrowband or wideband. If the time window is short, then its Fourier transform will be a wideband and a longer time returns a narrowband spectrogram [16]. The STFT general equation of a signal S is given by equation (1):

$$S(m,k)=\sum_{n=0}^{N-1} s\left(n+mN'\right)w(n)e^{-j\frac{2\pi}{N}nk}, \quad (1)$$

where

k=[0:K] is the kth Fourier coefficient.

K=N/2 is the frequency index corresponding to the Nyquist frequency.

S(m,k) indicates the m-index time-frequency(frame) spectrogram.

N=window segment length.

N'=the shifting step of the time window.

$w(n)$=windowing method of an N - point sequence.
N' should be smaller than N in order to produce an overlap between the time windows. S depends on the window function; in practice, different window shapes are used, such as: Symmetric, Unimodal and Gaussian.

The spectrogram contains a compromise between time resolution and frequency resolution: a large window provides less localization in time and more discrimination in frequency. The window obtains a time-slice of the signal, during which the spectral characteristics are nearly constant [16]; the obtained segments shift the time window with some overlapping. The spectrogram is defined as the magnitude of $S(m,k)$, represented as $A(m,k)$, as show in equation (2):

$$A(m,k)=\frac{1}{N}|S(m,k)|^2. \quad (2)$$

The spectrogram resolution can be enhanced modifying the length window; a long window provides a better frequency resolution, but poor time resolution. A short window, however, provides better time resolution but poor frequency resolution. A good visualization in the spectrogram depends the selection of an appropriate window length and overlapping. Fig. 2 shows a spectrogram of a signal, which is a time-varying spectral representation of a signal. A spectrogram layout is usually as follows: the x-axis represents time, the y-axis represents frequency, and the third dimension is amplitude (spectral content) of a frequency-time pair, which is color coded. This three dimensional data can also create a 3D plot, where the intensity is represented as height on the z-axis but a 2D chart provides a better understanding.

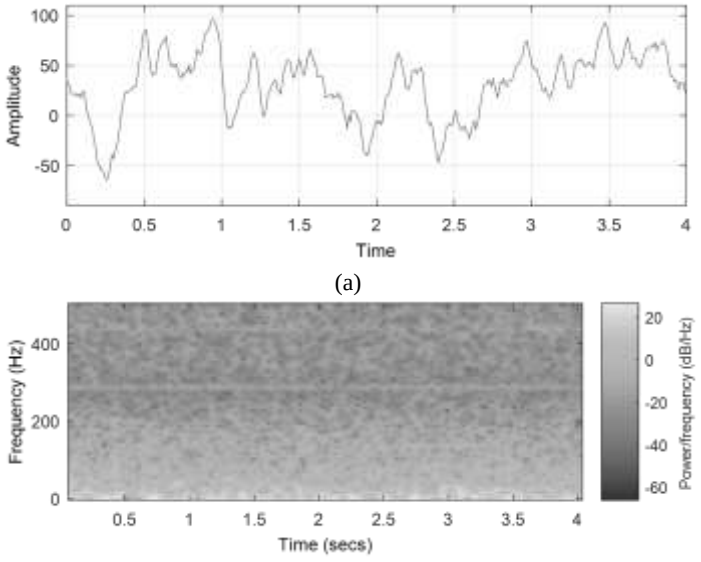

Fig. 2. a) EEG Signal with different amplitude along time, b) EEG signal Spectrogram.

2.2.1 Gray-Level Co-occurrence Matrix (GLCM)

GLCM is a statistical method for examining texture, which considers the spatial relationships between pixels. Statistical measures can be extracted with this method, such as contrast, correlation, and energy, among others. Different authors use these features and others for classification; in an analysis done by Mustafa et al., [2] to classify mental stages through spectrograms, the authors extracted 80 statistical features for four orientations of the matrix, reducing the features vector by applying PCA, and used K nearest neighbors(KNN) to classify the stages. In [18], a comparison was made of two classifiers —Support Vector Machine and Artificial Neural Networks— following the same methodology, but the features vector had other statistical features, improving the accuracy for KNN (using Euclidian distance) in comparison with the Artificial Neural Network (ANN). A BCI system [19] based on motor imagery acts in real-time using a single channel to classify the left- and right-hand motor imagery signals; these features are texture descriptors, employing a logistic regression classifier in offline mode.

A methodology for detecting two different datasets —epileptic and sleep stages— is in [20]. By extracting statistical features of GLCM, these features were encoded using the Fisher Vector (FV) to be applied in an extreme machine learning for two different tests —in sleep stages and in epileptic stages. Another method implemented in [21] for classifying epileptic seizure electroencephalogram (EEG) signals uses three texture descriptors to extract features: GLCM, the texture feature coding method (TFCM), and local binary patterns (LBP). GLCM and TFCM provided statistical features and LBP used the histogram; SVM was employed with LIBLINEAR in the classification with different accuracies, obtaining the highest accuracy with LBP and GLCM. Automatic person authentication [7] uses statistical features from spectrograms, but the energy obtained is used as a feature as well; a sum of the distance of the features is calculated to reduce dimension, and finally, ANN and SVM are applied.

In [22], co-occurrences of histograms of oriented gradients (CoHOG) and Eig(Hess)-CoHOG features are extracted from the spectrogram; the features obtained are used in a nonnegative least squares classifier (NNLS), to classify alcoholism and control EEG signals.

2.2.2 Frequency Bands

An EEG Signal is commonly described in terms of brain activities; these are divided into frequency bands, which have a certain biological significance and different properties. Delta (δ) has the highest amplitude and the slowest wave; it is associated with deep sleep and waking states. Theta (θ) has an amplitude greater than 20 μV and a range of 4-7 Hz, and is linked with idling, creative inspiration, unconscious material, drowsiness, and deep meditation. Alpha (α) has an amplitude of 30-50 μV and a range of 8-13 Hz, it is usually associated with relaxation, concentration, and sometimes with attention. Mu (μ), is found in the alpha wave and is regularly related to suppression, indicating that the motor neurons are working. Beta (β) is linked to alertness, thinking, and active concentration, and falls in the range between 12 and 30 Hz. Lastly, Gamma

(γ) with a frequency greater than 30 Hz, is seen during short term memory matching of recognized objects, sounds, or tactile sensations [8].

The previous section showed different techniques for extracting features from a spectrogram based on GLCM. Some authors have obtained the spectrograms from a specific band or from several bands; however, features are extracted directly from GLCM. In this section, methods for extracting features based on spectrograms of different frequency bands are shown. A method based on these spectrograms [15] for comparing obtained components during writing and imagined writing is implemented; to obtain the spectrogram, the STFT is applied to EEG signals, and the resulting spectrograms are modified to have components only in the frequency domain. All components are compared employing a correlation, concluding that the writing and imagined writing features are the same.

The classification of an epileptic seizure [23] employs the extraction of Frequency Cepstral Coefficients and stochastic relevance analyses on spectrograms of different rhythms. SVM is applied, obtaining competitive accuracies for different datasets. On the other hand, [24] extracts Malmquist-Takenaka coefficients from Spectrograms and statistical features, with the same objective. In this methodology, STFT is in discrete form, and the classification uses an Alternating Decision Tree (ADTree) classifier with three different datasets.

Another work [25] identifies a motor imagery database of left- and right-hand movements, applying multivariate empirical mode decomposition (MEMD) to generate multiple intrinsic mode functions (IMFs) to the EEG recordings. STFT is applied to the most significant mode. Features such as peak and entropy of the magnitude spectrum are used for the KNN classifier. Another work that implements KNN is [26], using Kd-Trees in this case; the features are the energy of the spectrograms in different frequency bands, obtaining a different accuracy and varying the number of features; the aim of this work is to detect drowsiness automatically.

3 Discussion

The reviewed literature in previous sections shows methods implemented for analyzing EEG signals based on spectrograms, using STFT as a representation in time-frequency, due to the competitive results compared with analyses based on time or frequency domain, as well as representations. This section shows relevant characteristics of the techniques in EEG processing. Table 2 shows the analysis of different methodologies, with attributes such as authors (reference), type of dataset to classify (objective), a brief description of the data (information), (#) the number of classes to use, and finally, three columns (preprocessing, feature extraction, and classification) to identify the techniques used in the analysis phases.

3.1 EEG Analysis

In section two, the steps to follow in the EEG analysis were presented. The first step removes artifacts and noise; a few experiments [19,23,25] apply low-pass and band-

pass filters, which can be implemented easily, as the band frequencies are known. Most experiments do not apply this phase, although the datasets were acquired from the experiments. As general information, the following was identified: the number of classes (column #) that predominate for the analysis are only two, employing a maximum of three classes. Regardless of the approaches for the data sets, the method used for the preprocessing stage is filtering, which is applied in few experiments.

Table 2. Methods about EEG signals analysis based on spectrograms.

Reference	Objective	Information	#	Preprocessing	Feature extraction	Classification
Mustafa et al. 2010.	Image Classification	Proposed Dataset	3	-	Statistical features and PCA.	KNN (70.83%).
Mustafa et al. 2012.	Brain balancing classification	Proposed Dataset	3	-	Statistical features and PCA.	90% KNN Euclidian distance) and 87.5% ANN.
Duque, et al. 2014	Epileptic seizures.	100 Signals, single channel, 5 datasets, 173.61Hz, Bonn University.	2	Low-pass	Stochastic analysis and cepstrals coefficients.	95.78%-100%, SVM.
Kovács, et al, 2014.	Epileptic seizures.	100 Signals, single channel, 5 datasets, 173.61Hz, Bonn University.	2	-	Malmquist-Takenaka coefficients and statistical features.	96.7%, 98.36%, and 99.7%. alternating decision tree
Khairul, et al, 2015	Imagery motor.	1 person, 2 channel, 128 Hz.	2	Band-pass	Threshold.	KNN (90%)
Kumar and Sharma, 2015	Epileptic seizures.	100 Signals, single channel, 5 datasets, 173.61Hz, Bonn University.	2	-	Energy and PCA.	ANN(995%, 99.33% and 92.36%)
Nieves and Manian, 2016.	Authentication.	Easycap, 32 channels. C3,CZ,C4. Imagery movements.	2	-	Energy and statistical features. Sum of distances.	ANN(6-channels)-98%. SVM(2-3, channels)-90%, 96%.
Alcin, et al, 2016	Multiclass (sleep stages and Epileptic seizures)	100 Signals, single channel, 5 datasets, 173.61Hz, Bonn University. Proposed dataset, 100 Hz, 30s, 8 subjects.	2	-	GLCM y VF.	Extreme machine learning. 95.17%,95.38% - sleep stages. 96.40%-Epilepsy.
Camacho, et al. 2016.	Imagery motors in a BCI.	1 channel, 400Hz, 20 subjects.	2	Band-pass	GLCM. Statistical features.	Linear regression, 87.6%.
Jalilifard, et al., 2016	Drowsiness	10 subjects.	2	-	Energy statistical features. Random Forest.	88.57%-91% (KNN-Kd Trees)
Bajaj, et al.,2016	Alcoholism and mental control.	120 files, 64 electrodes, 256 Hz.	2	-	Gray scale, Co-HOG and Eig(Hess)-CoHOG Statistics.	91.67%, 91.67% and 95.83% (NNLS)
Segür, et al., 2016	Epileptic seizures.	100 Signals, single channel, 5 datasets, 173.61Hz, Bonn University. Proposed dataset, 100 Hz, 30s, 8 subjects.	2	-	GLCM, statistical features; LBP y TFCM.	100%, SVM and LIBLINEAR.

For the feature extraction, statistical values of the time-frequency representation are utilized; it was seen that feature selection methods were applied in only three experiments; some authors implemented this stage through PCA [2,3,27] [3], sum of distances

[7], and random forest [26]. Finally, the classification was generally performed by KNN, SVM, and ANN; SVM was the best classifier in regard to accuracy.

The methods described are focused on medical applications related to sleep, aspects of authentication, the detection of imaginary motor images, and others. Medical applications are aimed at the detection of brain abnormalities within the EEG signals, such as neurological disorders caused by epileptic seizures, and the effects of alcohol on the brain. In sleep analysis, experiments are performed for the detection of sleep states to identify problems of drowsiness. However, applications aimed at aspects of authentication, show results that can be competitive compared to other biometrics, demonstrating that the EEG signal has the sufficient properties to make it safe, unique, non-visible, and non-modifiable.

Table 3. Features extracted from spectrograms.

Reference	Feature extraction method	Features Obtained
Mustafa et al. 2010.	Statistical features and PCA.	Proposed features by Haralick and Soh: Autocorrelation, Contrast, Correlation, Cluster prominence, Cluster shade, Dissimilarity, Energy, Entropy, Homogeneity, Maximum probability, Variance, Sum average, Sum variance, Sum entropy, Different variance, Different entropy, Information of correlation 1, Information of correlation 2, Inverse difference normalized, and Inverse difference moment normalized.
Mustafa et al. 2012.	Statistical features and PCA.	Proposed features by Haralick, Soh and Clausi: Autocorrelation, Contrast, Correlation, Cluster prominence, Cluster shade, Dissimilarity, Energy, Entropy, Homogeneity, Maximum probability, Variance, Sum average, Sum variance, Sum entropy, Different variance, Entropy difference, Information of correlation 1, Information of correlation 2, Inverse difference normalized, and Inverse difference moment normalized.
Duque, et al. 2014	Stochastic analysis and cepstral coefficients.	Five Cepstral coefficients in frequency related to rhythms: alpha, beta, theta and delta.
Kovács, et al, 2014.	Malmquist-Takenaka coefficients and statistical features.	Features using the discrete STFT as from rational coefficients and five statistical values of the coefficients: Mean median, maximum, minimum, and standard deviation.
Khairul, et al, 2015	Threshold.	Peak and entropy of the magnitude spectrum.
Kumar and Sharma, 2015	Energy and PCA.	Fractional energy for EEG segments.
Nieves y Manian, 2016.	Energy and statistical features. Sum of distances.	Energy, mean, variance and skewness from spectrogram.
Alcin, et al, 2016	GLCM y VF.	Texture descriptors: Contrast, correlation, energy, and homogeneity.
Camacho, et al. 2016.	GLCM. Statistical features.	Texture descriptor: Correlation, energy, contrast, homogeneity, and dissimilarity.
Jalilifard, et al., 2016	Energy statistical features. Random Forest	The mean value of power in time and standard deviation (SD) and Shanon entropy related to each time-segment were computed from the time-domain.
Bajaj, et al.,2016	Gray scale, CoHOG and Eig(Hess)-CoHOG Statistics.	Features obtained from local texture: Co-occurrence of Oriented Gradient (CoHOG) and Eig(Hess).
Segür, et al., 2016	GLCM, Statistical features, LPB and TFCM.	Texture descriptors: Contrast, correlation, energy, and homogeneity.

From the different, it was noted that for the analysis of epilepsy diseases, there is a set of data that different authors [3,20,23,24] employ, for example use a dataset created by the University of Bonn with five different subsets (A-E), with 100 single channel

signals, a duration of 26.3 seconds and a sampling frequency of 173.61 Hz. This database contains high quality information (number of studies, time and number of seizures), which were validated by experts. Table 2 shows four works where this dataset was used; although only one implemented the preprocessing stage [23] and the accuracy obtained was between 95.78% and 100%.

3.2 STFT and Features

Different objectives are proposed in the analyzed works, such as classification of motor images, the detection of imagined words, and others; nevertheless, they do not show execution time, which could be due to the real-time application execution not being required. STFT implementation is fast, but resolution depends on the window selected. For the purposes of the experiments in the literature, the resolution obtained with STFT is enough, implying less computational work and less execution time. However, it was shown in [2,4,5] that dimensionality is a problem to be solved; experiments where PCA was used do not report execution time and [22] do not apply reduction of characteristics, although they do note that it is necessary to reduce the vectors, as they have thousands of values.

All works use different features, most of them based on texture; some works extract features from GLCM applied to spectrograms; although the features are similar, changing some of these affect accuracy. Table 3 shows features extracted from spectrograms by the reviewed analysis; the first column corresponds to the authors, the second shows the methods implemented for extraction, and the last presents the features obtained.

These works generally extract features from texture or from statistics based on spectrograms, while others use energy as their main feature and employ different techniques, such as LBP, TFCM, CoHOG, and Cepstral coefficients, among others. Different methods are implemented; however, the most widely used is based on texture descriptors and energy.

4 Conclusions

The Short-Time Fourier Transform provides a quick implementation for different approaches; acceptable results have been reported in terms of classification, derived from an extraction of spectrogram features obtained by applying STFT to the EEG signal. However, despite the results in the literature with this representation, there is a limited scope, considering the complexity of the characteristics, which presents different research challenges, such as computational performance, accuracy in the classification, data dimension reduction, filtering of noisy signals, and others. Therefore, it is necessary to propose improved methodologies that can meet these challenges in all phases of analysis.

In some experiments, spectrograms have shown low resolution, which implies a changing performance; however, a high resolution requires more computational work and time. To solve this, a different purpose of feature extraction can be proposed, in addition to a feature selection method, because of the large dimensional data vector.

Due the information provided by spectrograms, it is possible to apply a classification based on thresholds. The literature on this methodology does not usually explain how the threshold is obtained, but another way to solve this, could be through the implementation of genetic algorithms.

Most of the methodologies here presented analyze a single dataset as a seizure, a dream, alcoholism, imaginary movements, or others, but only one of these at a time; however, a different methodology could be proposed to classify more than one EEG signal. [20] classifies two datasets with competitive results in terms of accuracy, compared with methodologies whose results come from one analysis to a single dataset.

References

1. Ridouh, A., Boutana, D., Benidir, M.: Comparative Study of Time Frequency Analysis Application on Abnormal EEG Signals. Recent Advances in Electrical Engineering and Control Applications, Springer International Publishing, Cham 355–368 (2017)
2. Mustafa, M., Taib, M., Murat, Z., Sulaiman, N., Aris, S.: The Analysis of EEG Spectrogram Image for Brainwave Balancing Application Using ANN. In: 2011 UkSim 13th International Conference on Computer Modelling and Simulation, 64–68 (2011)
3. Kumar, J., Bhuvaneswari, P.: Analysis of Electroencephalography (EEG) Signals and Its Categorization-A Study. Procedia Engineering, 38, 2525–2536 (2012)
4. Ilyas, M., Saad, P., Ahmad, M.: A survey of analysis and classification of EEG signals for brain-computer interfaces. In: 2015 2nd International Conference on Biomedical Engineering (ICoBE), 1–6 (2015)
5. Gandhi, V.: Chapter 2 - Interfacing Brain and Machine. Brain-Computer Interfacing for Assistive Robotics, Academic Press, San 7–63 (2015)
6. Grossi, E., Olivieri, C., Buscema, M.: Diagnosis of autism through EEG\ processed by advanced computational algorithms: A pilot study. Computer Methods and Programs in Biomedicine, 142, 73–79 (2017)
7. Nieves, O., Manian, V.: Automatic person authentication using fewer channel EEG motor imagery. In: 2016 World Automation Congress (WAC), 1–6 (2016)
8. Ramadan, R., Vasilakos, A.: Brain computer interface: control signals review. Neurocomputing, 223, 26–44 (2017)
9. Alamdari, N., Haider, A., Arefin, R., Verma, A., Tavakolian, K., Fazel-Rezai, R.: A review of methods and applications of brain computer interface systems. In: 2016 IEEE International Conference on Electro Information Technology (EIT), 345–350 (2016)
10. Al-Ani, A., Koprinska, I., Naik, G., Khushaba, R.: A dynamic channel selection algorithm for the classification of EEG and EMG data. In: 2016 International Joint Conference on Neural Networks (IJCNN), 4076–4081 (2016)
11. Ren, W., Han, M., Wang, J., Wang, D., Li, T.: Efficient feature extraction framework for EEG signals classification. In: 2016 Seventh International Conference on Intelligent Control and Information Processing (ICICIP), 167–172 (2016)
12. Li, Y., Wang, X., Luo, L., Li, K., Yang, X., Guo, Q.: Epileptic Seizure Classification of EEGs using Time-Frequency Analysis based Multiscale Radial Basis Functions. IEEE Journal of Biomedical and Health Informatics 1–12 (2017)
13. Wang, Y., Xu, G., Zhang, S., Luo, A., Li, M., Han, C.: EEG\ signal co-channel interference suppression based on image dimensionality reduction and permutation entropy. Signal Processing, 134, 113–122 (2017)

14. Lakshmi, M., Prakash, C.: Survey on EEG Signal Processing Methods (2014)
15. Zabidi, A., Mansor, W., Lee, Y., Fadzal, C.: Short-time Fourier Transform analysis of EEG signal generated during imagined writing. In: 2012 International Conference on System Engineering and Technology (ICSET), 1–4 (2012)
16. Boashash, B., ed.: Chapter 2 - Heuristic Formulation of Time-Frequency Distributions. Time-Frequency Signal Analysis and Processing, Second Edition, Academic Press, Oxford, 65–102 (2016)
17. El-Shennawy, K.: Communication Theory and Signal Processing for Transform Coding. Bentham Science Publishers (2014)
18. Mustafa, M., Taib, M., Murat, Z., Sulaiman, N.: Comparison between KNN and ANN Classification in Brain Balancing Application via Spectrogram Image. Journal of Computer Science & Computational Mathematics, 2(4), 17–22 (2012)
19. Camacho, J., Manian, V.: Real-time single channel EEG motor imagery based Brain Computer Interface. In: 2016 World Automation Congress (WAC), 1–6 (2016)
20. Alçin, Ö., Siuly, S., Bajaj, V., Guo, Y., Şengu¨r, A., Zhang, Y.: Multi-category EEG\ signal classification developing time-frequency texture features based Fisher Vector encoding method. Neurocomputing, 218, 251–258 (2016)
21. Şengür, A., Guo, Y., Akbulut, Y.: Time--frequency texture descriptors of EEG signals for efficient detection of epileptic seizure. Brain Informatics, 3, 101–108 (2016)
22. Bajaj, V., Guo, Y., Sengur, A., Siuly, S., Alcin, O.: A hybrid method based on time-frequency images for classification of alcohol and control EEG signals. Neural Computing and Applications, 1–7 (2016)
23. Duque-Muñoz, L., Espinosa-Oviedo, J., Castellanos-Dominguez, C.: Identification and monitoring of brain activity based on stochastic relevance analysis of short-time EEG rhythms. BioMedical Engineering OnLine, 13, 123 (2014)
24. Kovacs, P., Samiee, K., Gabbouj, M.: On application of rational Discrete Short Time Fourier Transform in epileptic seizure classification. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5839–5843 (2014)
25. Bashar, S., Hassan, A., Bhuiyan, M.: Motor imagery movements classification using multivariate EMD and short time Fourier transform. In: 2015 Annual IEEE India Conference (INDICON), 1–6 (2015)
26. Jalilifard, A., Pizzolato, E.: An efficient K-NN approach for automatic drowsiness detection using single-channel EEG recording. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 820–824 (2016)
27. Deng, L., OŚhaughnessy, D.: Speech Processing: A Dynamic and Optimization-Oriented Approach. Taylor & Francis (2003)

ISSN 1870-4069

Modelos para la generación automática de diálogos: Una revisión

Andrés Vázquez, David Pinto, Darnes Vilariño, Mauricio Castro

Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, Puebla, México

{andrex,dpinto,darnes}@cs.buap.mx

Resumen. La generación automática de diálogos es una parte importante en la interacción humano-robot, los diálogos generados deben de garantizar una conversación coherente entre el humano y el robot, se espera que la interacción sea lo más natural y eficaz que se pueda, es por eso por lo que en este trabajo se presenta una revisión de algunos modelos que se han utilizado para la generación de diálogos en la interacción humano-computadora.

Palabras clave: Sistema de diálogo, generación de lenguaje natural, interacción humano-robot, redes neuronales.

Models for Automatic Generation of Dialogues: A Review

Abstract. The automatic generation of dialogues is an important part of the human-robot interaction, the generated dialogues must guarantee a coherent conversation between the human and the robot, the interaction is expected to be as natural and effective as possible, that is why this paper presents a review of some models that have been used to generate dialogues in the human-computer interaction.

Keywords. Dialogue system, generation of natural language, human-robot interaction, neural networks.

1. Introducción

A medida que mejora la comprensión del lenguaje y la tecnología de generación automática de diálogos, existe un interés creciente en la construcción de sistemas de conversación de usuario, que pueden ser utilizados para una variedad de aplicaciones tales como planificación de viajes, sistemas tutoriales o soporte técnico basado en chat.

La mayoría de los trabajos en este contexto han enfatizado la comprensión o la generación de una secuencia de palabras asociadas con una sola oración o turno de un orador, potencialmente aprovechando el turno anterior. Más allá del contexto local, el uso del lenguaje en una conversación orientada a objetivos refleja el tema global de discusión, así como el rol respectivo de cada participante.

En los sistemas de diálogo hablado (SDS Spoken Dialogue Systems), la tarea de generación de lenguaje natural (NLG Natutal Language Generation) es convertir una representación con significado (MR Meaning Representation) producida por el gestor de diálogo en una o más oraciones en un lenguaje natural.

En este trabajo se realiza el análisis de algunos métodos utilizados en la generación de diálogos, con el objetivo de tener un punto de referencia en este contexto y en el futuro poder realizar trabajos que permitan proponer nuevas técnicas o mejoras en los modelos que ya existen de generación de diálogos.

2. Sistemas de diálogo

Los sistemas de diálogo hablado o sistemas conversacionales (SDS) son una tecnología concebida para facilitar la interacción natural mediante el habla entre una persona y una computadora. Constituyen como una interfaz o un intermediario entre un usuario y un sistema de cómputo, que tiene la ventaja de no requerir el uso de una pantalla, un teclado o un ratón y de recurrir, en cambio, al medio de comunicación propio de los seres humanos.

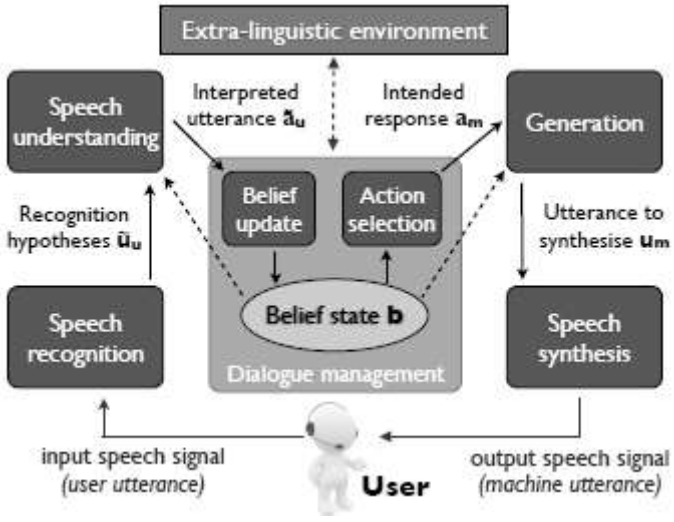

Fig. 1. Esquema de la arquitectura del sistema de diálogo, Fuente Lison [2].

Un sistema de diálogo consiste de una estructura modular en la que cada módulo se ocupa de unas determinadas tareas en interacción con todos los demás. La figura 1 describe una arquitectura de un sistema de diálogo, el cual consta de un módulo de

reconocimiento de habla que se ocupa de transformar la señal sonora en una representación que pueda procesar el módulo de análisis semántico; una vez realizado este análisis, interviene el módulo de gestión del diálogo, conectado a una interfaz que permite la consulta a una base de datos y que envía la información recogida al módulo de generación de respuestas; este último, con la intervención de un conversor de texto en habla, facilitaría al usuario la información deseada. Como elemento central se incluye el contexto del diálogo, que contiene información sobre el enunciado que se está procesando, la historia del diálogo y el estado actual de la interacción.

Puede observarse cómo cada uno de los módulos de un sistema de diálogo está asociado a uno o más modelos que reflejan los conocimientos necesarios para garantizar la interacción en el usuario y una aplicación computacional.

3. Algunas técnicas utilizadas en la generación de diálogos

Dentro de los trabajos que tienen que ver con la generación de diálogos podemos citar el trabajo de Konstantopoulos [1], donde se propone una arquitectura novedosa para los sistemas de diálogo afectivo y multimodal que permite un control explícito sobre los rasgos de personalidad que se quiere que el sistema exhiba. Más específicamente, acercarse a la personalidad como un medio de sintetizar diferentes y posiblemente conflictivos modelos de adaptación en un modelo general que se utilizará para impulsar los componentes de interacción del sistema. Además, esta síntesis se realiza en presencia de conocimiento de dominio, de manera que la estructura del dominio y las relaciones influyen en el cálculo de los resultados.

3.1. Modelos Bayesianos probabilísticos

En esta sección se describen dos trabajos relevantes con el enfoque bayesiano, el trabajo de Lison [2] muestra cómo representar la estructura subyacente de modelos probabilísticos para el diálogo utilizando reglas probabilísticas. Estas reglas se definen como asignaciones estructuradas sobre variables del estado de diálogo, especificadas usando condiciones y efectos de alto nivel. Estas reglas pueden incluir parámetros tales como probabilidades de efecto o utilidades de acción. Las reglas probabilísticas permiten al diseñador del sistema explotar poderosas generalizaciones en la especificación del dominio de diálogo sin sacrificar la naturaleza probabilística del modelo. El marco es muy general y puede expresar un amplio espectro de modelos, desde modelos clásicos totalmente estimados a partir de datos a aquellos que incorporan un rico conocimiento previo. La elección del modelo dentro de este espectro es, por lo tanto, esencialmente una decisión de diseño dependiente de las disponibilidades relativas de datos de entrenamiento y conocimiento de dominio. También se presentan algoritmos para construir redes bayesianas correspondientes a la aplicación de las reglas y para estimar sus parámetros a partir de datos utilizando la inferencia bayesiana como se muestra en la figura 2. El enfoque presentado se ha implementado en un sistema de diálogo oral para la interacción humano-robot y se ha validado en una tarea de aprendizaje de políticas basada en un conjunto de datos de Wizard-of-Oz.

Los resultados empíricos han demostrado que la estructura de reglas permite que el algoritmo de aprendizaje converja más rápido y con mejor rendimiento de generalización.

De igual forma en [3] presenta una evaluación de un sistema de diálogo oral que detecta y se adapta a la desconexión del usuario y a la incertidumbre en tiempo real. Los autores comparan esta versión de su sistema con una versión que se adapta sólo a la desconexión del usuario y a una versión que ignora por completo la desconexión del usuario y la incertidumbre. Encuentran un aumento significativo en el éxito de la tarea al comparar ambas versiones de adaptación de su sistema a su línea base no adaptativa, pero sólo para los usuarios masculinos. Su evaluación examina el impacto de la adaptación a diferentes números de estados afectivos en el éxito de la tarea, y también examina las interacciones con el género del usuario. Sin embargo, esta evaluación sólo se llevó a cabo en un escenario Wizard-of-Oz, donde un humano oculto reemplazó el reconocimiento de voz, el análisis semántico y los componentes de detección de afecto de su sistema de diálogo.

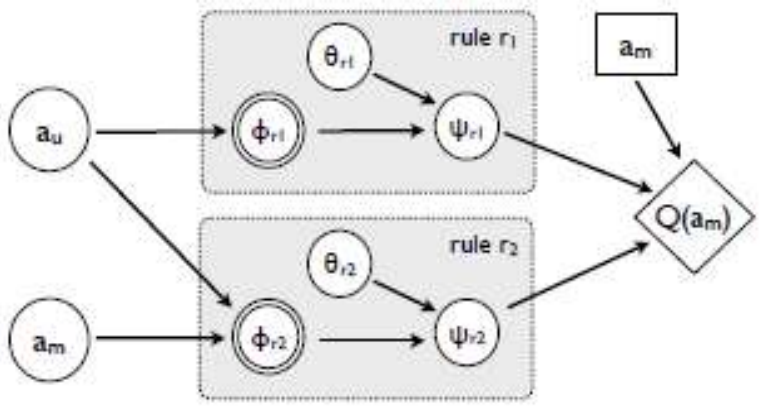

Fig. 2. Red bayesiana con las reglas r_1 y r_2, fuente Lison [2].

3.2. Modelos estocásticos

En esta sección se describen algunos trabajos que utilizan el modelo basado en procesos de decisión de markov, tal es el caso de Barlier [4] que propone un marco original para modelar los diálogos hablados de humano-computadora para tratar la coadaptación entre usuarios y sistemas de diálogos hablados en tareas no cooperativas. La conversación es modelada como un juego estocástico: tanto el usuario como el sistema tienen sus propias preferencias, pero tienen que llegar a un acuerdo para resolver una tarea no cooperativa. Ellos son entrenados conjuntamente para que el gestor del diálogo aprenda la estrategia óptima contra el mejor usuario posible. Los resultados obtenidos por simulación muestran que se aprenden estrategias no triviales y que este marco es adecuado para el modelado del diálogo.

Ahora bien, en [5] se plantea que, en los últimos años, con la difusión de los robots domésticos, la necesidad de mejorar las capacidades de comunicación de esos robots

con las personas ha ido en aumento. El objetivo de este estudio es construir un marco para un sistema de diálogo que se ocupe de la información multimodal que un robot observa. Ellos aplican el proceso de decisión de Markov parcialmente observable (POMDP por sus siglas en inglés) para modelar la interacción multimodal entre un ser humano y un robot. A través de sus experimentos, han confirmado que el marco propuesto funciona correctamente y logran una interacción multimodal eficaz con un robot.

En los trabajos [6,7] utilizan el aprendizaje de refuerzo (RL) para aprender las políticas de diálogo de preguntas y respuestas para una aplicación del mundo real. En el trabajo de Misu y Georgila [6] se analiza un corpus de interacciones de visitantes del museo con dos personajes virtuales que sirven como guías en el Museo de Ciencias de Boston, con el fin de construir un modelo realista de comportamiento del usuario al interactuar con estos personajes. Se construye un usuario simulado sobre la base de este modelo y se utiliza para aprender la política de diálogo de los personajes virtuales usando RL. En este trabajo se sigue un enfoque basado en Procesos de Decisión de Markov Parcialmente Observable (POMDP).

En el trabajo de Lison [8] las contribuciones son dobles. En primer lugar, demostrar cómo aplicar el RL Bayesiano basado en modelos para aprender el modelo de transición de un dominio de diálogo. También se comparan dos enfoques de modelado en el contexto de un escenario humano-robot en el que se instruye a un robot Nao a moverse y recoger objetos. Los resultados empíricos muestran que el uso de representaciones estructuradas permite al algoritmo de aprendizaje converger más rápido y con mejor desempeño de generalización. La idea clave es modelar la gestión del diálogo como un proceso de decisión de Markov (MDP) o un proceso de toma de decisiones de Markov parcialmente observable (POMDP), y dejar que el sistema aprenda por sí mismo la mejor acción a realizar en cada posible situación conversacional mediante interacciones repetidas con un usuario (real o simulado). En [9, 10, 11] también se utiliza aprendizaje de refuerzo como se muestra en la figura 3 donde el modelo de diálogo y la política están parametrizados y dada una función de recompensa adecuada, pueden optimizarse utilizando el aprendizaje de refuerzo. Estudios empíricos han demostrado que las políticas optimizadas a través de RL son generalmente más robustas, flexibles y adaptables que sus homólogos hechos a mano.

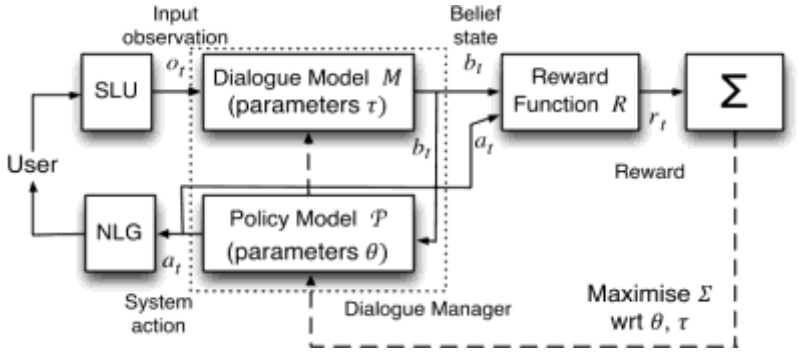

Fig. 3. Componentes de un sistema de diálogo hablado basado en POMDP, fuente Young [11].

3.2.1. Proceso de Decisión de Markov

Un Proceso de Decisión de Markov MDP [7] es formalmente una tupla (S, A, R, T, γ) donde S es el espacio de estado, A es el espacio de acción, $R: S \to \Re$ es la función de recompensa, $T: S \times A \to P(S)$ es un conjunto de probabilidades de transición Markoviana y γ es un factor de descuento $0 \leq \gamma \leq 1$. La optimización del problema de la toma de decisiones consiste en encontrar una política $\pi: S \to P(A)$ que mapea estados a acciones de tal manera que las recompensas acumulativas obtenidas siguiendo esta política se maximiza. Para ello, la calidad de una política se mide en cada estado como la recompensa acumulativa esperada que puede obtenerse siguiendo la política que comienza en este estado. Esta medida se denomina función de valor $V^{\pi}: S \to \Re$:

$$V^{\pi}(s) = E\left[\sum_{i=0}^{\infty} \gamma^i R(s_i) | s_0 = s, a_i = \pi(s_i)\right]. \quad (1)$$

Se puede definir un orden sobre funciones de valor tales como $V^1 > V^2$ si $\forall s\ V^1(s) > V^2(s)$. La política óptima π^*es el que maximiza la función de valor para cada estado: $\pi^* = \arg max_{\pi} V^{\pi}$.

3.2.2. Proceso de Decisión de Markov Parcialmente Observable

Un POMDP [6] se define como una tupla (S, A, P, R, O, Z, γ, b_0) donde S es el conjunto de estados (que representan contextos diferentes) en los que el sistema puede estar (el mundo del sistema), A es el conjunto de acciones del sistema, $P: S \times A \to P(S, A)$ es el conjunto de probabilidades de transición entre estados después de tomar una acción, $R: S \times A \to \Re$ es la función de recompensa, O es un conjunto de observaciones que el sistema puede recibir sobre el mundo, Z es un conjunto de las probabilidades de observación $Z: S \times A \to Z(S, A)$ y γ es un factor de descuento que pondera las recompensas a largo plazo. En cualquier momento dado el paso i el mundo está en algún estado no observado $s_i \in S$. Porque si no se conoce exactamente se mantiene una distribución sobre estados llamados estado de creencia b, así $b(s_i)$ es la probabilidad de estar en estado s_i, con estado de creencia inicial b_0. Cuando el sistema realiza una acción $\alpha_i \in A$ basada en b, siguiendo una política $\pi: S \to A$, recibe una recompensa $r_i(s_i, \alpha_i) \in \Re$ y la transición al estado s_{i+1} de acuerdo con $P(s_{i+1}|s_i, \alpha_i) \in P$. El sistema recibe entonces una observación o_{i+1} según $P(o_{i+1}|s_{i+1}, \alpha_i)$. La calidad de la política π seguido por el agente se mide por la recompensa futura esperada también llamada Q-función, $Q^{\pi}: S \times A \to \Re$.

3.3. Modelos de redes neuronales

En los trabajos [12, 13] presentan un nuevo sistema de generación de lenguaje natural para sistemas de diálogo hablado capaces de incorporar (adaptar) la forma de hablar de los usuarios, proporcionando respuestas contextualmente apropiadas. El generador se basa en redes neuronales recurrentes (RNN) y el enfoque de secuencia a secuencia. Es completamente entrenable a partir de datos que incluyen el contexto anterior junto con las respuestas que se generarán. Muestran que el generador de con-

texto genera mejoras significativas sobre la línea base tanto en métricas automáticas como en una prueba de preferencia por parejas.

En [14] se presenta una arquitectura de red neuronal para la generación de respuestas que son a la vez sensibles al contexto y basadas en datos. Como tal, puede ser entrenada de extremo a extremo en cantidades masivas de datos de medios sociales. Comentan que, esta es la primera aplicación de un modelo de red neuronal a la generación de respuesta de dominio abierto, y creen que el presente trabajo sentará las bases para modelos más complejos por venir. Además, introducen una nueva técnica de extracción de múltiples referencias que muestra una promesa para la evaluación automatizada.

Así también en [15] se presenta un enfoque simple del modelado conversacional. El modelo conversa al predecir la oración siguiente dada la oración u oraciones anteriores en una conversación. La fuerza de este modelo es que puede ser entrenado de extremo a extremo y por tanto, requiere mucho menos reglas hechas a mano. Además, este modelo sencillo puede generar conversaciones sencillas dado un gran conjunto de datos de entrenamiento de conversación. Los resultados preliminares sugieren que, a pesar de optimizar la función objetivo errónea, el modelo es capaz de conversar bien. Es capaz de extraer conocimiento de un conjunto de datos de dominio específico y de un conjunto de datos grande, ruidoso y de dominio general de subtítulos de películas. En un conjunto de datos ruidoso de transcripción de películas de dominio abierto, el modelo puede realizar formas sencillas de razonamiento de sentido común.

En [16] proponen un enfoque basado en redes neuronales que modela los procesos de atención e intención. Consiste esencialmente en tres redes neuronales recurrentes (RNN). La red de codificadores es un modelo de nivel de palabra que representa oraciones origen. La red de intenciones es una red recurrente que modela la dinámica del proceso de intención. La red decodificadora es una red recurrente que produce respuestas a la entrada desde el origen. Es un modelo de lenguaje que depende de la intención y tiene un mecanismo de atención para atender a las palabras de origen en particular, al predecir un símbolo en la respuesta. El modelo está entrenado de extremó a extremó sin datos etiquetados. Los experimentos muestran que este modelo genera respuestas naturales a las entradas de los usuarios. En la figura 4 se describe el modelo atención con intención y en la tabla 1 se da un ejemplo del proceso de diálogo.

Tabla 1. Un ejemplo de proceso de diálogo [16].

user	*computer is infected*
agent	do you want to retrieve the files that was deleted?
user	the ones that the virus deleted, yes.
agent	i can help you resolve the issue with our virus removal and protection service
user	ok
agent	here is a link how to run system restore
user	thank you.
agent	you welcome

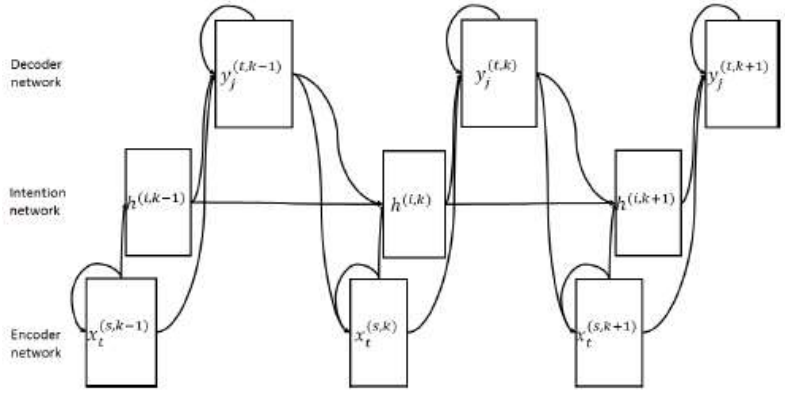

Fig. 4. El modelo de atención con intención (AWI). El modelo se desenrolla en tres turnos. En cada turno tiene RNNs para la red del codificador y la red del decodificador. Cada sesión está representada por un vector de dimensión fija, que es un estado oculto de una red de intención RNN, Fuente Yao [16].

En [17] se presenta un generador de lenguaje estadístico basado en una estructura de memoria a corto plazo (LSTM por sus siglas en inglés) controlada semánticamente. El generador de LSTM puede aprender de datos no alineados optimizando conjuntamente sus componentes de planificación de frases utilizando un criterio de entrenamiento de entropía cruzada simple sin ninguna heurística y se obtiene una variación de calidad de lenguaje simplemente mediante muestreo aleatorio de las salidas de la red. Con menos heurística, una evaluación objetiva en dos diferentes dominios de prueba muestran que el método propuesto mejoró el rendimiento en comparación con los métodos anteriores. Los jueces humanos anotaron el sistema LSTM más alto en informatividad y naturalidad y en general lo prefirieron a los otros sistemas.

Utilizando el mismo enfoque de estructura de memoria (LSTM) [18] presenta un modelo de conversación que incorpora tanto el contexto como el rol de participante en las conversaciones de dos partes. Se exploran diferentes arquitecturas para integrar la información de roles y contextos de los participantes en un modelo de lenguaje de memoria a corto plazo (LSTM). El modelo conversacional puede funcionar como un modelo de lenguaje o un modelo de generación de lenguaje. Los experimentos realizados en el corpus de diálogo de Ubuntu muestran que el modelo puede capturar la interacción de múltiples turnos entre los participantes. El método propuesto supera a un modelo LSTM tradicional medido por la perplejidad del modelo de lenguaje y la clasificación de respuestas. Las respuestas generadas muestran diferencias de características entre los dos roles de los participantes.

En [19] se comenta que los modelos de red neuronal de secuencia a secuencia para la generación de respuestas conversacionales tienden a generar respuestas habituales y seguras (por ejemplo, no sé) independientemente de la entrada. Sugieren que la función objetivo tradicional, es decir, la probabilidad de salida (respuesta) dada entrada (mensaje) no es adecuada para tareas de generación de respuesta. En su lugar, proponen utilizar Máxima Información Mutua (MMI por sus siglas en inglés) como la fun-

ción objetivo en los modelos neuronales. Los resultados experimentales demuestran que los modelos de MMI propuestos producen respuestas más diversas, interesantes y apropiadas, produciendo ganancias sustanciales en las puntuaciones de BLEU4 en dos conjuntos de datos conversacionales y en evaluaciones humanas.

Tabla 2. Análisis comparativo de los diferentes modelos utilizados en la generación automática de diálogos.

Modelo	***Autor***	***Herramientas, técnicas y/o modelos***	***Dominio***
Reglas Probabilísticas	Lison (2012)	Redes Bayesianas	Conjunto de datos de izard-of-Oz.
	Ritter (2013)	**SMT**, GIZA++	Twitter, WMT08, WMT09
Procesos estocásticos	Barlier et al., (2015)	MDP, POMDP y RL	100000,75000 diálogos
	Iijima y Kobayashi (2016)	POMDP, Q-learning	
	Lison (2013)	POMDP, Redes Bayesianas	Wizard-of-Oz y simulación de usuario
	Pietquin (2013)	MDP, IRL	Simulación del usuario
	Misu et al., (2012)	MDP, POMDP, RL	Corpus (Museo de Ciencias de Boston)
	Png (2011)	MDP, POMDP, BRL	Corpus SACTI1
	Mahadik (2017)	POMDP, Redes Bayesianas	
	Young y Thomson (2013)	POMDP y RL	
Redes Neuronales	Dušek y Jurčíček (2016)	LSTM y seq2seq	Crowdsourcing
	Sordoni et al., (2015)	RNN, RLM	Conversaciones de Twitter
	Vinyals (2015)	RNN (LSTM) y seq2seq	OpenSubtitles (dominio abierto). Conjunto de datos de solución de problemas (dominio cerrado)
	Yao et al., (2015)	RNN, seq-seq	Diálogos de servicio de chat
	Wen et al., (2015)	SC-LSTM	Ontologías (SF Restaurant SF Hotel)
	Luan et al., (2016)	LSTM	Corpus de diálogo UBUNTU
	Li et al., (2016)	NN, seq-seq y MMI	OpenSubtitles
	Su et al., (2016)	NN, SL y RL	Conjunto de datos de Wizard-of-Oz (Amazon Mechanical Turk)
	Li et al., (2016)	LSTM y seq2seq, RL	OpenSubtitles
	Shang et al., (2015)	NN, **SMT, GIZA++**	Microblog similar a Twitter en China
	Sutskever (2014)	RNN, LSTM, seq-seq	WMT'14

En la investigación [20] se describe un enfoque de dos pasos para la gestión del diálogo en sistemas de diálogo hablado orientados a tareas. Se propone un marco de red neuronal unificado para permitir al sistema aprender primero mediante supervisión a partir de un conjunto de datos de diálogo y luego mejorar continuamente su comportamiento a través del aprendizaje por refuerzo, todos utilizando algoritmos basados en gradiente en un solo modelo. Los experimentos demuestran la efectividad del modelo supervisado en la evaluación basada en el corpus, con simulación de usuarios y con sujetos humanos remunerados. El uso del aprendizaje por refuerzo mejora aún más el rendimiento del modelo en ambos entornos interactivos, especialmente bajo condiciones de mayor ruido.

Los recientes modelos neuronales de generación del diálogo [14, 16, 17, 18, 19, 21] ofrecen una gran promesa para generar respuestas para los agentes de conversación, pero tienden a ser miopes, prediciendo enunciados uno a la vez ignorando su influencia en los resultados futuros. Modelar la dirección futura de un diálogo es crucial para generar diálogos coherentes e interesantes, una necesidad que llevó a los modelos tradicionales de procesamiento de lenguaje natural (PNL por sus siglas en ingles) de diálogo a basarse en el aprendizaje por refuerzo. Como en el trabajo [22] que muestra cómo integrar estos objetivos, aplicando el aprendizaje por refuerzo profundo para modelar la recompensa futura en el diálogo chat-bot. El modelo simula diálogos entre dos agentes virtuales, utilizando métodos de políticas de gradiente para recompensar las secuencias [23] que muestran tres propiedades conversacionales útiles: informatividad, coherencia y facilidad de respuesta.

En la tabla 2 se hace un análisis de los modelos, autores, herramientas, técnicas y el dominio de datos empleados en la generación automática de diálogos de los artículos analizados en este trabajo.

4. Conclusión

En este trabajo se presentó una revisión de diferentes modelos para la generación de diálogos desde los modelos Bayesianos probabilísticos, los estocásticos hasta los modelos basados en redes neuronales, como los modelos que utilizan LSTM secuencia-secuencia (SEQ2SEQ) que son un tipo de modelo de generación neuronal que maximiza la probabilidad de generar una respuesta dado el turno anterior del diálogo y actualmente los modelos que buscan integrar SEQ2SEQ y los paradigmas de aprendizaje por refuerzo, aprovechando las ventajas de ambos. Esto da pauta para futuras investigaciones en donde se puedan generar respuestas cada vez más diversas e interactivas que fomenten una conversación más sostenida.

Agradecimientos. Queremos agradecer a Vicerrectoría de Investigación y Estudios de Posgrado de la Benemérita Universidad Autónoma de Puebla por apoyar este trabajo a través del proyecto Extracción de frases clave y sus relaciones en artículos científicos.

Referencias

1. Konstantopoulos, S.: An Embodied Dialogue System with Personality and Emotions. In: Proceedings of the 2010 Workshop on Companionable Dialogue Systems, ACL, 31–36 (2010)
2. Lison, P.: Probabilistic Dialogue Models with Prior Domain Knowledge. Association for Computational Linguistics. In: Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Seoul, South Korea, 179–188 (2012)
3. Litman, D., Forbes-Riley, K.: Evaluating a Spoken Dialogue System that Detects and Adapts to User Affective states. In: Proceedings of the SIGDIAL 2014 Conference, Philadelphia, U.S.A., ACL, 181–185 (2014)
4. Barlier, M., Perolat, J., Laroche, R., Pietquin, O.: Human-Machine Dialogue as a Stochastic Game. In: Proceedings of the SIGDIAL 2015 Conference, Prague, Czech Republic, 2–11 (2015)
5. Iijima, S., Kobayashi, I.: A POMDP-based Multimodal Interaction System Using a Humanoid Robot. In: 30th Pacific Asia Conference on Language, Information and Computation (PACLIC 30) Seoul, Republic of Korea, 519–525 (2016)
6. Misu, T., Georgila, K., Leuski, A., Traum, D.: Reinforcement Learning of Question-Answering Dialogue Policies for Virtual Museum Guides. In: Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Seoul, South Korea, 84–93 (2012)
7. Pietquin, O.: Inverse Reinforcement Learning for Interactive Systems. ACM, 978-1-4503-2019-1/13/08 MLIS'13, Beijing, China (2013)
8. Lison, P.: Model-based Bayesian Reinforcement Learning for Dialogue Management. arXiv:1304.1819v1 [cs.AI] (2013)
9. Png, S., Pineau, J.: Bayesian Reinforcement Learning for POMDP-Based Dialogue Systems. In: Conference Paper in Acoustics, Speech, and Signal Processing, 1988, ICASSP-88., 1988 International Conference on 2011 (2011). DOI: 10.1109/ICASSP.2011.5946754
10. Mahadik, S., Dwivedi, P., King, M., Zhu B., Tang S.: Dialogue Manager for Spoken Dialogue System: Review. In: International Conference on Emanations in Modern Technology and Engineering (ICEMTE-2017), 27–29 (2017)
11. Young, S., Gasˇic´, M., Thomson, B., Williams, J. D.: POMDP-Based Statistical Spoken Dialog Systems: A Review. In: Proceedings of the IEEE, 101(5) (2013)
12. Dušek, O., Jurčíček, F.: A Context-aware Natural Language Generator for Dialogue Systems. In: Proceedings of the SIGDIAL 2016 Conference, Los Angeles, USA, 185–190 (2016)
13. Dušek, O., Jurčíček, F. 2016b.: A context-aware natural language generation dataset for dialogue systems. In Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation. 6–9 (2016)
14. Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Dolan, B.: A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In: Proceedings of NAACL-HLT (2015)
15. Vinyals, O., Le Q.: A Neural Conversational Model. In: Proceedings of the 31st International Conference on Machine Learning, Lille, France, JMLR: W&CP, 37 (2015)
16. Yao, K., Zweig, G.: Attention with Intention for a Neural Network Conversation Model. In: Presented at NIPS Workshop on Machine Learning for Spoken Language Understanding and Interaction 2015, arXiv:1510.08565v3 [cs.NE] (2015)
17. Wen, T., Gašić, M., Mrkšić, N., Su, P., Vandyke D., Young, S.: Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In: Proceedings

of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 1711–1721 (2015)
18. Luan, Yi., Ji, Y., Ostendorf, M.: LSTM based Conversation Models. arXiv:1603.09457v1 [cs.CL] (2016)
19. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A Diversity-Promoting Objective Function for Neural Conversation Models. In: Proceedings of NAACL-HLT 2016, Association for Computational Linguistics, San Diego, California, 110–119 (2016)
20. Su, P., Gašić, M., Mrkšić, N., Rojas-Barahona, L., Ultes, S., Vandyke, D., Young S.: Continuously Learning Neural Dialogue Management. arXiv:1606.02689v1 [cs.CL] (2016)
21. Shang, L., Lu, Z., Li, H.: Neural Responding Machine for Short-Text Conversation. arXiv.1503.02364v2[cs.CL] (2015)
22. Yao, K., Zweig, G.: Attention with Intention for a Neural Network Conversation Model. In: Presented at NIPS Workshop on Machine Learning for Spoken Language Understanding and Interaction 2015. arXiv:1510.08565v3 [cs.NE] (2015)
23. Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., Jurafsky, D.: Deep Reinforcement Learning for Dialogue Generation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, 1192–1202 (2016)
24. Sutskever, I., Vinyals, O., Le, Q.: Sequence to Sequence Learning with Neural Networks.arXiv:1409.3215v3 [cs.CL] (2014)

ISSN 1870-4069

Implementation of Police Patrols Based on an Intelligent Model of VRP

María B. Bernábe[1], Alberto Ochoa Zezzatti[2], Rogelio González[1], Jorge Ruiz Vanoye[3], Martín Estrada Analco[1]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, Mexico

[2] Universidad Autónoma de Ciudad Juárez, Maestría en Cómputo Aplicado, Mexico

[3] Universidad Autónoma del Estado de Hidalgo, Mexico

{rgonzalez, mestrada}@cs.buap.mx, beatriz.bernabe@gmail.com, alberto.ochoa@uacj.mx

Abstract. In this paper we present the implementation of police patrols based on an intelligent model of VRP using an algorithm of variable neighborhood search, which allows to solve vehicle routing situations, obtaining results that show optimization, problems arising from transportation, distribution and logistics; in most markets, transport means a value added to goods, for which the use of computerized methods of transportation resulting in significant savings. Most optimization problems of the real world are dynamic, that is, because the information available about the situation you want to solve is variable over time. One example is the problem of routing of dynamic order in which is required to develop a service plan for a set of clients using a fleet of vehicles, in order to include in the plan to new customers who send their orders along the route or the workday. We use MATLAB for algorithm development, data visualization, and numeric computation. With ant colony algorithm to calculate the total time of the patrols.

Keywords. Police patrols, dynamic optimization, variable neighborhood search algorithm, vehicle routing problems, Matlab.

1 Introduction

The VRP constitutes a series of problems that can be formulated mathematically by means of directed graphs, otherwise the VRP is a method that aims to optimize the resources for the production of any type of product to the consumer or customer and services.

Today, technology tools can be used for troubleshooting using simulation systems or software, these systems are based on mathematical algorithms in a way users of these systems develop their logic [5,7].

Therefore, most optimization problems are dynamic, that is, form the information changes over time. In these cases, the algorithms that are able to adapt to a changing environment can provide greater results than a search reset after each change. An optimization problem in the real world of great interest is the vehicle routing problem (VRP). Described for the first time in [3], with applications in the field of transport and telecommunications. One goal of this problem is to reduce the cost of the routes taken by a fleet of vehicles to service the requests of a number of customers.

Currently, is possible to run a fleet of vehicles in real time, thanks largely to advances in the field of ICT, such as the use of sensors to measure traffic flow, global positioning systems, or GPS to determine the exact position of the vehicles in case of theft or to monitor execution paths for each car, mobile communication systems to provide real-time information, etc. In this content, you can define characteristics involving dynamic information in the classical VRP, namely, problems such as the shortest path between two clients can be blocked by an accident or because weather or also that customers change the routes of their orders, etc. Considering the above explained, the dynamic VRP are a number of different problems, very important in the industry and that they can carry out a study to improve efficiency in distribution systems [2,7].

The goal of most basic VRP is to deliver and collect goods for a set of customers with known demands, with minimal cost, finding optimal routes that begin and end in a reservoir where each client is visited once and the vehicles carrying load.

To make this work we consulted articles in PDF for VRP or GVRP, book, manual, examples and demos of Matlab, as well as the book design of experiments Roman Vaez and VRP thesis, paper related to the topic.

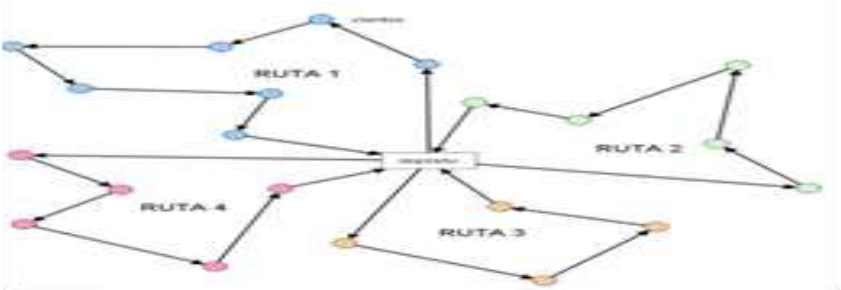

Fig. 1. Generic representation of a VRP Fuente [7].

A route is a simple cycle containing an origin and a destination and represents the sequence of visits made by a vehicle traveling the route, the cost and time of a route is obtained by adding the costs and times of the arcs that form the cycle.

1.1 Features VRP

According to [5] the main features of VRP are: The network transport is considered a terrestrial network, but can also be considered an airline or shipping network or a combination of both to problems that are seen in other types of transport. Customers: are characterized by a demand that must be satisfied by a vehicle and in many cases demand are assets that have a place in the vehicle. Deposits: vehicles are those that are responsible for distributing the goods for which should be directed to deposits and routes begin and end in the same tank. Routes: the objective is to minimize fixed and total costs, the number of vehicles per route, transportation time and distance as well as the benefits and customer satisfaction.

1.2 Types of VRP Problems

Problems with capacity constraints (PTRC) is a variant of VRP in which a fleet of vehicles to serve a number of clients from a tank at minimal cost, your goal is to minimize the fleet, the sum of the time routes and total demand for each customer.
Problems with time windows (VRPTM) are the same problem as the VRP with the difference that is required to serve customers in a given period of time. Its aim is to minimize the fleet, the amount of travel time and waiting time.

Problem with multiple deposits (MDVRP): when a company has several stores where you can serve your customers, if customers are around the deposits, the distribution may be modeled by a group of VRP's, however, if customers and the deposits are mixed, is different and its modeling is performed through a MDVRP.

Problems with delivery and returns (VRPPD) is a variant of VRP where a customer who has received a sent also have some merchandise that needs to be gathered by which it should be aware that products that customers sent must not exceed the vehicle capacity, this need hinders the problem of planning and exceeds the capacity of the vehicles, the distances or increases the need for a larger vehicle.

Problem of partial deliveries (SDVRP) is an advantage of generic VRP problem because it allows the same client to be visited by various delivery vehicles. This advantage is important if the size of the customer demand is as great as the capacity of the vehicles.

Random value problem (SVRP) is performed in two stages to reach a solution. The first is to determine before knowing the value of the variables and the second corrective action is taken when the values of the variables are known.

Periodic problem VRP (PVRP) is planning an extension of N days, your goal is to minimize the fleet and the total time of transport to serve all customers. During the course of N days each client must be visited at least once [7].

Matlab is an abbreviation of MATrix LABoratory because it supports vector and matrix operations that are fundamental to the solution of engineering problems and science, whatever with tools called toolboxes that extend graphical environment capacity to solve specific problems in specific areas.

MATLAB is a computer language that allows you to perform operations and faster than other languages like C, C ++ and FORTRAN computational tasks. The tasks are

the development of algorithms, data visualization and numerical computations and can be used in applications such as signal analysis and imaging, communications, control, test and measurement, analysis, financial modeling and computational biology [1].

1.3 Description of Nearest Neighbor Algorithm

The main idea of the algorithm of the nearest neighbor method is to make a number of copies of their behavior over time, namely, that the information on the latest series match the latest information available before observation t +1.

The purpose of Nearest Neighbor algorithm is to locate similar pieces of information, regardless of location in time [8].

1.4 Types of VRP Problems

Patrol or policing action is set displacement on routes established and controlled, exercising oversight in a given territory. And areas are performed by using a preventive police.

Procedure 1: Algorithm of a local search for the closest neighbor [4]

```
Generating first solution (X);
         Make
         X '= structure building (σ);
         If (f (X ') <f (X)) then;
         X '= best solution found;
         X = X '
        End = yes
 While end stop criterio
```

Within the zone or area to watch is: Urban area which consists of industrial area, residential area and commercial area. Surveillance routes for urban area are: neighborhoods, schools, household, temples, parks, flea markets, recreational facilities, etc.

One of the purposes of police patrols is to prevent, stop in fragrance, protect, and encourage citizen participation and support the administration of justice, in coordination with various police forces and government.

To meet the patrols of police is necessary to establish monitoring devices which are operational plans that help fulfill the purposes of police surveillance, operations were performed with the use of equipment and material to be alert [9].

1.5 The Patrols can be Classified into Five Types [9]

Stationary patrol: It is performed by an element that is responsible for granting security to a specific place such as: companies, shops, booths modules and security and surveillance which are supported by auto patrols or other mobile.

Mobile patrol: It is carried out by using a conveyance as may be patrols, bicycles, motorcycles, horses and even takes place on foot, to make the patrol certain areas are crossed with special attention to everything that is out of order.

Mixed patrol: This takes more than just a specific type of monitoring devices usually is the combination of mobile patrol or system closed-circuit monitoring.

Patrol monitored: They are made using instruments or electronic devices and video devices remotely controlled by one or two audio items.

With the use of technology, we can make a satellite patrol [9]: It is done through GPS or global positioning systems that allow us to find the exact location of a mobile anywhere in the world.

2 Methodology

MATLAB program for the nearest neighbor algorithm determines the number of patrols and total distance, source [11].

```
NumberOfPatrols=length (Patrols);
SetsOfNumberOfOutletsInPatrol= [];
for PatrolsIndex=1: NumberOfPatrols
    NumberOfOutletsInPatrol=length     (Patrols     {1,
    PatrolsIndex});
    SetsOfNumberOfOutletsInPatrol=
    [SetsOfNumberOfOutletsInPatrol
    NumberOfOutletsInPatrol];
end
ttt =find(SetsOfNumberOfOutletsInPatrol(:)<=2);
Patrols (ttt) = [];
n=length (Patrols);
for PatrolIndex=1:n
    Patrol_VRP= Patrols {1, PatrolIndex}
    r=length (Patrols_VRP);
    jum = 0;
    for t=1:r-1,
        subrute   =jum+   (d(Patrol_VRP(t),   Patrol_VRP
        (t+1)));
        jum=subrute;
    end
    DistanceSets (PatrolIndex,:)=[jum];
    TotalDistance=sum(DistanceSets);
end
NumberOfPatrols =n
TotalDistance=TotalDistance
img = imread ('map.jpg'); %<==File name of your map
min_x = 0;
```

```
  max_x = 200;
  min_y = 0;
  max_y = 200;
  x=Problem (:,2);
  y=Problem (:,3);
  figure
  x_min = min_x;
  x_max = max_x;
  y_min = min_y;
  y_max = max_y;
  imagesc ([x_min x_max ], [y_min y_max], img);
  %Colouring Line
  for tyt=1:n
      hold on
      shortestPath =Patrols{1,tyt};
      colour =mod(tyt,6);
      xd=[x(shortestPath)];
      yd=[y(shortestPath)];
      for i=2:length(shortestPath)-1
         text(xd(i),yd(i),['  Punto de revision ',num2str
        (shortestPath (i))]);
      end
      text (xd(1),yd(1),['Caseta de vigilancia ']);
      if colour==1
         plot(xd,yd,'-
cs','LineWidth',2,'MarkerEdgeColor','k',...
                  'MarkerFaceColor','g',...
                  'MarkerSize',10)
              plot(x(1),y(1),'MarkerEdgeColor','k',...
                  'MarkerFaceColor','k',...
                  'MarkerSize',10)
```

Table 1. Results of the execution of the request.

Police patrol	Points instance
Patrol_1_VRP	5, 14, 9, 10, 2
Patrol_2_VRP	7, 4, 11, 15, 24, 23
Patrol_3_VRP	22, 13, 21, 12, 3
Patrol_4_VRP	8, 20, 6. 19
Patrol_5_VRP	18, 17, 16, 25
Number a police patrol	5
Distance total	623.3026

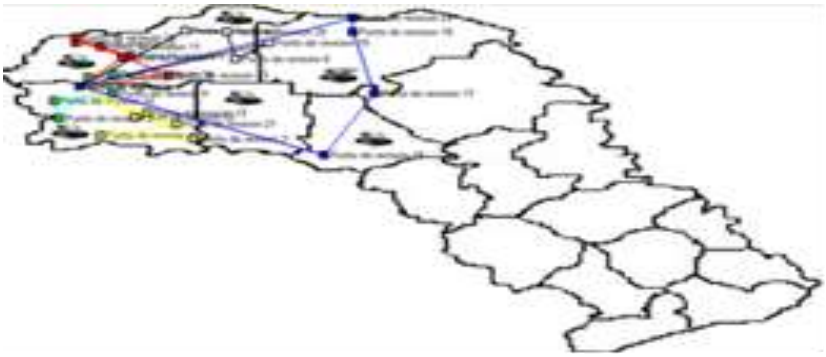

Fig. 2. Map with patrols of Chihuahua capital.

The map shows where population is located and where most criminal activity is located so it is necessary to go more than once for a police patrol.

The population density according to INEGI in 2010 in this capital city of Chihuahua is 14 people per square kilometer.

The test results of the ant colony code show that 60%, 80% and 100% of allocated vehicles can cover an area of 5 patrols police in a colony, as shown in Figure 4 which correspond to eastern Chihuahua.

C ++ program ant colony to determine the total travel time of the patrols. The printing area must be 122 mm × 193 mm. The text is justify for each paragraph.

```
typedef struct {
    int current_city;
    int next_city;
        unsigned int tabu[NUM_CITIES];
        int tour_index;
        unsigned int tour[NUM_CITIES];
        double tour_length;
} ANT_T;
CITY_T cities[NUM_CITIES];
ANT_T  ants[NUM_ANTS];
double pheromone[NUM_CITIES][NUM_CITIES];
double precomputed_distance[NUM_CITIES][NUM_CITIES];
int    best_index;
double best_tour = 100000.0;
void cargar(void){
FILE *fp;//apuntador de archivo
register int i;
int MAX=NUM_CITIES;
int x,y,z;
if((fp=fopen("att48.tsp", "r"))==NULL){
```

```
printf("No se puede abrir el archivo.\n");
return; }
```

2.1 Design of Experiments for the Ant Colony Program

25 colonies
60% of vehicles
Beta = 1
Rho = 0.9

Number of executions	*Iteration*	*Results*
1	1325	352.33
2	800	342.95
3	1025	343.99
4	250	344.40
5	300	352.33
6	275	356.51
7	600	369.84
8	1225	353.67
9	525	349.21
10	375	353.03

25 colonies
60% of vehicles
Beta = 1
Rho = 0.95

Number of executions	*Iteration*	*Results*
1	575	343.81
2	275	345.26
3	425	355.01
4	775	353.34
5	475	345.27
6	350	347.76
7	300	340.00
8	1550	368.31
9	375	338.56
10	1225	349.32

25 colonies
60% of vehicles
Beta = 5
Rho = 0.9

Number of executions	*Iteration*	*Results*
1	458	358.73
2	850	344.40
3	550	340.00
4	1300	348.35
5	1200	356.27
6	825	357.43
7	450	343.81
8	525	357.26
9	400	343.99
10	1200	355.77

25 colonies
60% of vehicles
Beta = 5
Rho = 0.95

Number of executions	*Iteration*	*Results*
1	325	339.41
2	475	345.26
3	400	368.31
4	650	345.26
5	625	343.99
6	550	364.05
7	1900	346.91
8	875	349.03
9	275	354.52
10	1150	345.43

25 colonies
80% of vehicles
Beta = 1
Rho = 0.9

Number of executions	*Iteration*	*Results*
1	300	353.79
2	900	345.26
3	350	363.29
4	400	353.08
5	475	346.90
6	1359	349.21
7	1575	345.43
8	275	344.40
9	950	355.81
10	250	343.81

25 colonies
80% of vehicles
Beta = 1
Rho = 0.95

Number of executions	*Iteration*	*Results*
1	275	348.27
2	1775	342.95
3	1550	356.17
4	200	343.81
5	925	339.41
6	450	342.95
7	875	345.26
8	2075	348.62
9	400	362.47
10	700	340.00

25 colonies
80% of vehicles
Beta = 5
Rho = 0.9

Number of executions	*Iteration*	*Results*
1	225	346.77
2	775	349.21
3	900	340.86
4	175	343.81
5	575	347.76
6	150	345.80
7	925	346.80
8	775	345.80
9	1075	339.41
10	275	346.91

25 colonies
80% of vehicles
Beta = 5
Rho = 0.95

Number of executions	*Iteration*	*Results*
1	125	340.80
2	400	338.56
3	1725	340.86
4	1475	340.00
5	225	347.77
6	1950	342.79
7	2325	348.13
8	2100	346.80
9	650	343.99
10	275	351.65

25 colonies
100% of vehicles
Beta = 1
Rho = 0.9

Número de ejecuciones	*Iteración*	*Resultados*
1	475	362.77
2	225	374.81
3	675	343.99
4	1250	340.00
5	175	348.62
6	700	353.49
7	1300	352.22
8	1725	361.92
9	725	353.77
10	900	347.76

25 colonies
100% of vehicles
Beta = 1
Rho = 0.95

Number of executions	*Iteration*	*Results*
1	325	348.35
2	1650	338.56
3	1350	353.03
4	150	359.36
5	275	343.81
6	1875	374.43
7	1125	350.09
8	2050	346.91
9	1325	361.92
10	475	343.99

25 colonies
100% of vehicles
Beta = 5
Rho = 0.9

Number of executions	*Iteration*	*Results*
1	650	348.13
2	825	352.33
3	1825	347.76
4	625	340.00
5	275	343.99
6	1600	339.41
7	550	352.33
8	850	345.80
9	2350	340.86
10	1600	345.43

25 colonies
100% of vehicles
Beta = 5
Rho = 0.95

Number of executions	*Iteration*	*Results*
1	675	340.00
2	1675	347.06
3	175	339.41
4	1575	345.43
5	124	351.23
6	2425	350.10
7	1350	352.33
8	675	347.06
9	1900	348.27
10	1100	340.86

% of vehicles	*Beta*	*Rho*	*Iteration*	*Results*
60	1	0.9	800	342.95
60	1	0.95	375	338.56
60	5	0.9	550	340.00
60	5	0.95	325	339.41
80	1	0.9	250	343.81
80	1	0.95	925	339.41
80	5	0.9	1075	339.41
80	5	0.95	400	338.56
100	1	0.9	1250	340.00
100	1	0.95	1650	338.56
100	5	0.9	1600	339.41
100	5	0.95	175	339.41

Table 2. Results of experimental design.

% of vehicles	Beta	Rho	Iteration
60	1	0.9	800
60	1	0.95	375
60	5	0.9	550
60	5	0.95	325
80	1	0.9	250
80	1	0.95	925
80	5	0.9	1075
80	5	0.95	400
100	1	0.9	1250
100	1	0.95	1650
100	5	0.9	1600
100	5	0.95	175

The first column shows the percentage of ants.
Vehicle of 60% = 15.
Vehicle of 80% =20.
Vehicle of 100% = 25.

These results were obtained with 60%, 80% and 100% of the assigned vehicles can cover 5 patrols, a colony.

Based on the results shown in the table it was obtained that the best values for beta, and rho is the percentage of vehicles:
Beta = 1.
Rho = 0.95.
% Vehicles = 100%

In order to be able similar, the most efficient arrangement of individuals in a social network, we developed an atmosphere able to store the data of each one of the representing individuals of each society, this with the purpose of distributing of an optimal form to each one of the evaluated societies. One of the most interesting characteristics observed in this experiment was the diversity of the cultural patterns established by each community. The scenes structured associated with the agents cannot be reproduced in general, since they only represent a little while dice in the space and time of the different societies. These represent a unique form and innovating of adaptive behavior which solves a computational problem that it does not try to clustering the societies only with a factor associated with his external appearance (attributes of each society), trying to solve a computational problem that involves a complex change between the existing relations. The generated configurations can be metaphorically related to the knowledge of the behavior of the community with respect to an optimization problem (to select culturally 47 similar societies, without being of the same quadrant [3]).

Table 3. Orthogonal array.

Variable Value								
A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**Color**
H	H	H	H	H	H	H	L	1
H	H	H	H	H	H	L	H	2
H	H	H	H	H	L	H	H	3
H	H	H	H	L	H	H	H	3
...	...	...	...	...	...	...	...	...

The main experiment consisted of detailing each one of the 1087 communities, with 500 agents, and one condition of unemployment of 50 époques, this allowed us to generate the best selection of each Quadrant and their possible location in a Diorama, which was obtained after comparing the different cultural and social similarities from each community, and to evaluate with Multiple Matching Model each one of them [10]. The developed tool classified each one of the societies pertaining to each quadrant, with different wardrobe for societies that included linguistic identity and for societies only with cultural identity; this permit identifies changes in the time respect at other societies.

The design of the experiment consists in an orthogonal array test, with the interactions between the variables: emotional control, ability to fight, intelligence, agility, force, resistance, social leadership, and speed. These variables are studied in a range of color (1 to 64).

The orthogonal array is L-N(2**8), in other words, 8 factors in N executions, N is defined by the combination of possible values of the 8 variables an the possible range of color (To see Table 1).

3 Conclusions and Future Work

In this paper the different VRP problems were analyzed and using the nearest neighbor algorithm map with police patrols which shows what will be the patrol largest number of routes are generated.

Using ant colony algorithm principle to understand its operation and analyzing the structure of the code for operation at program execution. It is concluded that 60%, 80% and 100% of vehicles can cover a land area of 5 patrols.

Future work of this research is the implementation of algorithms for solving such problems route, aimed at companies to improve their product delivery logistics, to improve service quality, timely delivery and satisfaction of clients.

References

1. Báez, D. et al.: MATLAB Con Aplicaciones a la Ingeniería, Física y Finanzas. 2a edición, 4–287 (2016)
2. Carrasco, R. et al.: TIC para una logística más sostenible (2015)
3. Dantzig G. B. and Ramser J. H.: The truck dispatching Problem. Operations Research Management Sciences, 6(1), 80–91 (1959)
4. Bentz, J.L., Kozak, J.J.: Influence of geometry on light harvesting in dendrimeric systems. II. Nth-nearest neighbor effects and the onset of percolation. Journal of luminescence, 62–74 (2006)
5. Ponce. J. et al.: Algoritmo de colonia de hormigas en CUDA para la optimización de rutas de distribución. Komputer Sapiens by Published
6. Rodas, J. et al.: Use of GVRP as a model of two specific real world problems and its bioinspired solution. Chapter by published. The book, Handbook of Research on Military, Aeronautical, and Maritime Logistics and Operations
7. Rodríguez, J.: Caracterización, Modelado y Determinación de las Rutas de la Flota en una Empresa de Rendering e-reading by Published
8. Sarasola, B. et al.: Un algoritmo de búsqueda en vecindario variable para la asignación de rutas a vehículos con pedidos dinámicos. SIMD by Published.
9. Técnicas de la intervención policial, programa de homologación en formación inicial para policías municipales
10. Inegi: Nombre de la página www.cuentame.inegi.org.mx. Densidad de población en Chihuahua capital. Consulted 2017
11. WahidS.H. Mathworks. Home page http://www.mathworks.com/matlabcentral/news reader/view_thread/324824. Programa en Matlab del vecino más cercano. Consulted 2015

Impreso en los Talleres Gráficos
de la Dirección de Publicaciones
del Instituto Politécnico Nacional
Tresguerras 27, Centro Histórico, México, D.F.
noviembre de 2017
Printing 500 / Edición 500 ejemplares