DSIE'13
Doctoral Symposium in
Informatics Engineering

Proceedings
of the
8th Doctoral Symposium
in
Informatics Engineering

24th and 25th January 2013, Porto, Portugal

Editors:
A. Augusto Sousa
Eugénio Oliveira

www.fe.up.pt/dsie13

FEUP EDIÇÕES

Sponsors

U.PORTO FEUP Faculdade de Engenharia Universidade do Porto
DEI Departamento de Engenharia Informática
aefeup
weDO technologies
MEGAtrónica
BRAGASOFT Informática
maselga

COPYRIGHT

1ª Edição/ 1st Edition 2013

ISBN: 978-972-752-151-7

Editors: A. Augusto Sousa and Eugénio Oliveira

Faculdade de Engenharia da Universidade do Porto

Rua Dr. Roberto Frias, 4200-465 Porto

DSIE'13 SECRETARIAT:

Faculdade de Engenharia da Universidade do Porto

Rua Dr. Roberto Frias, s/n

4200-465 Porto, Portugal

Telephone: +351 22 508 21 34

Fax: +351 22 508 14 43

E-mail: dsie13@fe.up.pt

Symposium Website: http://www.fe.up.pt/dsie13

FOREWORD

2013 Doctoral Symposium in Informatics Engineering - DSIE'13 is an event representing the 8th edition of a scientific meeting usually organized by PhD students of the FEUP Doctoral Program in Informatics Engineering (ProDEI). However, the current edition is co-organized with PhD students of MAP-tele, Doctoral Program in Telecommunications. This synergy brought up expectations on both quantity and quality increase of the submitted papers.

DSIE meetings have been held since the school year 2005/06 and the main goal has always been to provide a forum for discussion on, and demonstration of, the practical application of a variety of scientific research issues, particularly in the context of information technology, computer science, computer engineering plus, in the current edition, telecommunications. DSIE symposium comes out as a natural conclusion of mandatory ProDEI and, this year, MAP-Tele courses called, respectively, "Methodologies for Scientific Research" (MSR) and "Seminar", leading to a formal evaluation of the students learned competencies.

The aim of those specific courses (MSR and Seminar) is to give students the opportunity to learn the processes, methodologies and best practices related to scientific research, particularly in the referred areas, as well as to improve their own capability to produce adequate scientific texts. With a mixed format based on multidisciplinary seminars and tutorials, the course culminates with the realization of DSIE meeting, seen as a kind of laboratory test for the concepts learned by students. In the scope of DSIE, students are expected to play various roles, such as authors of the articles, both scientific and organization committee members as well as reviewers, duly guided by more senior lecturers and professors.

DSIE event is then seen as a "leitmotif" for the students to write scientific correct and adequate papers following the methods and good practices currently associated to outstanding research activities in the area. Although, still at an embryonic stage, and despite some of the papers still lack of maturity or mainly report a state of the art, we already can find some interesting research work or interesting perspectives about future work. At

this time, it was not essential, nor even possible, for most of the students in the first year of their PhD, to produce strong and deep research results. However, we hope that the basic requirements for publishing an acceptable scientific paper have been fulfilled.

DSIE'13 Proceedings include 17 articles accepted in the previously defined context. They cover a large spectrum of topics in informatics, computer science, telecommunication and engineering areas and can be grouped according to six main clusters. These clusters include some different, although related topics, since, as it was expected, the paper themes are of a large diversity. These clusters are named as follows: Wireless Communications and Computer Networks (4 papers), Distributed Systems and Software Engineering (3 papers), Information Systems and Interoperability (3 papers), Machine Learning and Artificial Intelligence (3 papers), Computer Graphics and Image Processing (2 papers), Microwave and Circuit Design (2 papers).

The complete DSIE'13 meeting encompasses a two days program that includes also two invited talks by outstanding researchers in advanced Models for the Media public service and Music Data Processing.

Professors responsible for both ProDEI and MAP-tele programs current edition, are proud to participate in DSIE'13 meeting and would like to acknowledge all the students who have been deeply involved in the success of this event that, hopefully, will contribute for a better understanding of the themes that have been addressed during the above referred courses, the best scientific research methods and the good practices for writing scientific papers and conveying novel ideas.

Eugénio Oliveira and Augusto Sousa (ProDEI)
Henrique Salgado and Aníbal Ferreira (MAP-tele)
January 2013

FOREWORD – ORGANIZING AND SCIENTIFIC COMMITTEES

The chairs of DSIE'13 Organizing and Scientific Committees warmly welcome you to the 8th edition of the Doctoral Symposium in Informatics Engineering. We accepted the invitation to be a part of these committees with great honor. Organizing an event such as this one proved to be both a challenging and a complex task, of which all of us certainly derive great value.

The joint effort of all colleagues of the Doctoral Program in Informatics Engineering and the Doctoral Program in Telecommunications was instrumental in making this event a success. We anticipate that this effort is reflected in the quality of the communications and the organization.

We would like to thank all the senior members of the Scientific Committee for their involvement. We would also like to thank the significant collaboration of Vítor Carvalho (CICA) and Sandra Reis (DEI) of the Faculty of Engineering of the University of Porto.

And, above all, we thank you for being a part of DSIE'13!

Pedro Strecht and Erico Leão (Organizing Committee Chairs)
Roberta Barbosa, Nelson Rodrigues and Filipe Teixeira (Scientific Committee Chairs)

CONFERENCE COMMITTEES

STEERING COMMITTEE

A. Augusto Sousa
Eugénio Oliveira

ORGANIZING COMMITTEE CO-CHAIRS

Erico Leão
Pedro Strecht

ORGANIZING COMMITTEE

Alexandre Perez
Diego Jesus
Eduardo Pinto
Erico Leão
Fábio Pinto
Joel Gonçalves
Margarida Gomes
Nelson Rodrigues
Pedro Strecht
Roberta Oliveira
Tiago Carvalho

SCIENTIFIC COMMITTEE CO-CHAIRS

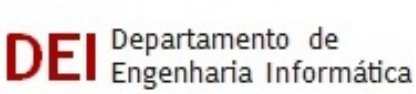
Filipe Teixeira
Nelson Rodrigues
Roberta Oliveira

SCIENTIFIC COMMITTEE

Adão Silva
Adriano Moreira
Ana Paiva
Ana Paula Rocha
Aníbal Ferreira
António Coelho
Carlos Soares
Eduarda Mendes Rodrigues
Gabriel David
Henrique Lopes Cardoso
Henrique Salgado
João Correia Lopes
João Manuel Tavares
João Mendes Moreira
João Paiva Cardoso
João Pascoal Faria
João Tiago Jacob
José Barbosa
José Machado da Silva
José Magalhães Cruz
José Ruela
Luís Alves
Manuel Ricardo
Ricardo Morla
Rosaldo Rossetti
Rui Aguiar
Rui Campos
Rui Maranhão
Rui Rodrigues
Tânia Calçada
Alex Araujo
Alexandre Perez
Bilal Hussain
Diego Jesus
Eduardo Pinto
Erico Leão
Fábio Pinto
Filipe Teixeira
Iman Kianpour
Joel Gonçalves
José Quevedo
Margarida Gomes
Nelson Rodrigues
Oluyomi Aboderin
Pedro Strecht
Roberta Oliveira
Syed Saqlain Ali
Tiago Carvalho
Zafeiris Kokkinogenis

SPONSORS

DSIE'13 – Doctoral Symposium in Informatics Engineering is sponsored by:



CONTENTS

TECHNICAL PROGRAMME

INVITED SPEAKERS

Artur Pimenta Alves
The Evolution of the Public Service in Broadcasting and the need of a new Ecosystem to support the collaboration of Public Operators, Companies and Citizens

Fabien Gouyon
Processing music data

INVITED SPEAKER

ARTUR PIMENTA ALVES

Artur Pimenta Alves is Full Professor at the Faculty of Engineering of the University of Porto, Portugal, was responsible for many teaching and research activities in the area of Telecom and, more recently digital media. He took part in the creation of INESC Porto, a research institute associated to that university, where he was member of the board and responsible for work in his areas of interest. He was responsible for the preparation of several European research projects under ESPRIT, RACE, Eureka, RACE, IST as well as many national research and development projects. He was responsible for the set up of an important collaboration with BBC in the area of networked digital TV production. Several SME's resulted from the activities developed at INESC in these projects. He created and directed for nearly 15 years the national committee that represented Portugal in the standardisation activities of MPEG. In recent years he integrated the direction of the digital media component of the program of collaboration between Portugal and the University of Texas in Austin and he created and was the director of a multidisciplinary PhD program in digital media developed under that program. Currently on leave from the university he is working for RTP, national Radio and TV Broadcaster as director of the Production Centre in Porto.

Talk:
The Evolution of the Public Service in Broadcasting and the need of a new Ecosystem to support the collaboration of Public Operators, Companies and Citizens

Abstract:
The presentation will start with a brief description of the evolution from the original model based in monopolist public service broadcasters to the present, characterising the changes in regulation and in particular the changes that resulted from the world digitisation and globalisation accelerated since the end of the 90's. A new method to analyse the possible scenarios for the

implementation of what is now called Public Service Media, based in a layered model will be described.
As a result of this process it is now generally accepted that a new ecosystem involving public operators, other publicly funded organisations, private companies and citizens will be necessary to support the development of the named "Digital Commons". This ecosystem will create new opportunities for creative companies certainly including those needed to develop the supporting technologies and new Apps and services.

INVITED SPEAKER

FABIEN GOUYON

Fabien Gouyon (PhD Computer Science, UPF Barcelona; MSc IRCAM Paris; MSc Signal Processing, ENSEEIHT Toulouse; BSc Theoretical Physics, UPS Toulouse) is Invited Assistant Professor at the Faculty of Engineering of the University of Porto, in Portugal, and senior research scientist and co-leader of the Sound and Music Computing Group of INESC TEC in Porto. His main research and teaching activities are in Music Information Retrieval and Music Pattern Recognition. He has published over 60 highly-cited papers in peer-reviewed international conferences and journals, published a book on computational rhythm description, gave the first tutorial on the topic at the International Conference on Music Information Retrieval in 2006 and participated to the writing of the European Roadmap for Sound and Music Computing, published in 2007. He was General Chair of the Sound and Music Computing Conference 2009 and General Chair of the International Society for Music Information Retrieval Conference 2012.

Talk:
Processing music data

Abstract:
Computer science and informatics engineering deal to a large extentwith the handling and processing of data.In this talk, I will focus on the processing of a particularlychallenging type of data: music. I will provide examples of the manydiverse modalities of music and illustrate the correspondingchallenges and opportunities faced in the rapidly-growing Music Information Retrieval field. I will finish by demonstrating somerecent research done at the Sound and Music Computing Group at INESCTEC (http://smc.inescporto.pt).

SESSION 1

COMPUTER GRAPHICS AND IMAGE PROCESSING

Using Geospatial Data for Procedural Urban Modelling

Diego Jesus
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias s/n
diego.jesus@fe.up.pt
António Coelho
INESC TEC, Departamento de Engenharia Informática,
Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias s/n
acoelho@fe.up.pt

***Abstract*—The urban procedural modelling is an expanding area in Computer Graphics. The techniques presented in this field allow the semiautomatic generation of virtual city spaces mainly through the use of textual rules and randomness. Such methods, however, are more adequate to the generation of fictional environments. To be able to reproduce real spaces, it is necessary to feed the techniques with large amounts of data in order to reduce the need for randomness. Such data is often stored in Geographic Information Systems (GIS) but the integration with procedural tools in an easy and uniform way still remains a problem. This paper presents a pipeline for the integration of both geometric and semantic data from multiple GIS data sources into procedural modelling techniques used for the generation of 3D virtual urban environments. Using such pipeline we were able to obtain a real city virtual model from georreferenced data in a certain level of visual fidelity.**

Keywords*—Procedural Modelling, GIS, Virtual Urban Environments

I. Introduction

The use of existing urban spaces in Computer Graphics applications has been increasing. However, modelling such spaces using traditional techniques requires too much effort and costs, which is a great downside in the development of any application. On the other hand, the procedural modelling techniques allow the semi-automatic generation of vast areas of urban environments, reducing the time and money spent on such process.

Research in this area has provided good results in the generation of fictional urban environments. There are still some issues, however, when applying these methods to generate spaces from real world scenarios. The reason behind this is that procedural modelling requires vast amounts of semantic data to produce environments that resemble closely enough the real ones. The most common work around for such lack of information is to introduce some randomness into the process, which in turn may result in an unpredictable set of models that may look quite different from the reality.

On the other hand, municipalities often store information regarding their urban space in Geographic Information Systems (GIS). This data can be gathered and processed in order to recreate existing urban spaces. Procedural modelling tools (such as CityEngine [1]) may allow the integration of such data but not in an uniform way. This means that data collected from different sources are fed into the system in a different way. In turn, it is then necessary to create procedural rules to accommodate each new data source. Also, if there is a need to create an urban area containing elements belonging to distinct municipalities, we may need to have production rules to generate similar elements from very different inputs.

In order to provide a more structured solution, the proposed pipeline first maps the original data into a Unified Model. Since different municipalities may have distinct amounts of information for each feature, there could be a need to amplify the existing data. That way, all elements that are fed into the modelling processes mapped to an uniform amount of associated semantic data. Only then, all the normalized data is fed to the procedural modelling tool in order to generate models to represent the urban space.

This paper is organised as follows: section II makes a brief overview of related work and section III presents the proposed pipeline, including the Data Mapping, Data Conversion and Amplification and Procedural Modelling steps. Section IV presents an overview of the pipeline's distributed architecture and in section IV we discuss the results obtained. Finally section VI provides the Conclusions and Future Work.

II. Related Work

Procedural modelling of urban environments stems from the use of L-Systems used in the work of Parish et. al. [2]. This mathematical tool, introduced in [3], was originally designed to model the growth process of simple muticellular organisms such as algae and fungi and were later adapted to model plants [4]. This means that these are systems meant for the modelling of growth processes in open spaces. Buildings, however, tend to have more spatial restrictions like a bounding volume. Because of this, L-Systems were considered to be inadequate to produce building models [5], [6]. To solve this, split and control grammars were introduced [5] based in the concept of shape. The former subdivides the spaces in each derivation until terminal symbols are found, representing the geometry, while the latter influences the choice of production rules to apply in each derivation step. Later, the CGA (Computer

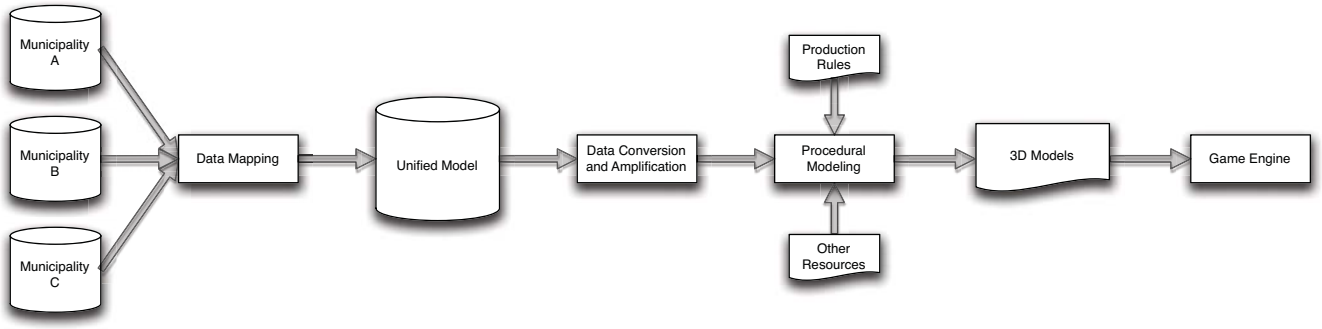


Fig. 1. Overview of the proposed Pipeline

Generated Architecture) Shape Grammar was introduced [6]. This grammar is capable of producing extensive architectural models with high detail. The implementation of CGA Grammar is integrated in the CityEngine [1] modelling tool.

One of the most common limitations in these systems is the reduced or nonexistence of spatial awareness, leaving out the possibility of performing queries to an object's surroundings. Geospatial L-Systems [7] were introduced to solve such limitation. These systems represent an extension to parametric L-Systems which incorporates spatial awareness. This approach combines the ability of data amplification provided by L-Systems with the geospatial awareness of GIS.

CityEngine provides the ability to create attribute layers that associate a value to a specific position in the terrain. In turn these values can be associated with production rules attributes to influence the appearance of the urban space. With these layers the user can control the road network generation, selections and properties of the models generated with CGA grammar rules [6]. However, the use of raster maps reduces the amount of semantic information that can be used. On the other hand, this tool is able to to import GIS files directly, which contain semantic data associated with geometry. The semantic data is then fed to the rules attributes. This is also the approach followed by Geospatial L-Systems [7]. This, however, requires the creation of specific rules for each case, or at least the modification of existing ones, leading to an increased effort.

To facilitate the integration of different GIS data sources in order to procedurally generate real urban spaces, an Urban Ontology was introduced in [8]. Since each different data source may contain different amounts of semantic data for the same type of element, the Level of Mapping (LOM) notion was also introduced. LOMs indicate the minimum required level of semantic information for each type of urban elements.

Trying to combine the power of procedural modelling techniques with Geographic Information Systems in order to produce virtual urban environments resembling real cityscapes led to the development of this work.

III. Pipeline for Geospatial Data

In this paper we propose a new pipeline for the generation of accurate 3D environments based on GIS data with the incorporation of semantic attributes. Such pipeline can be implemented in a distributed environment and can be integrated into any procedural modelling tool, given that they provide an interface allowing such integration. For this research, the chosen modelling tool was CityEngine [1]. The pipeline consists in the sequential execution of three processes: Data Mapping, Data Conversion and Amplification and Procedural Modelling. In Figure 1 we can see an overview of such pipeline and its several steps.

The first step, Data Mapping, is responsible for mapping the original GIS data into an unified data model based on the Urban Ontology presented in [8]. Such step is necessary due to the differences between data models belonging to different municipalities. Skipping this step would introduce vulnerabilities to such differences further down the pipeline, leading to the need of creating different processes for each new municipality or data source.

After the completion of the first step, the semantic and geometric information regarding the municipalities' urban elements is stored in the unified model. Nonetheless, the amount of information present in distinct databases can be quite different, meaning that the unified data may have elements in different Levels of Mapping [8]. To use only one set of procedural modelling processes in the next step, these should be provided with the same amount of information for every element, i.e., a specific LOM. As such the need to convert and amplify existing data arises.

The final step is responsible for the procedural modelling of the urban space itself and is conducted by CityEngine. This takes the geometric and semantic data provided by previous steps and generates three dimensional models representing the city environment. Another procedural modelling tool could be used as long as it provides a means of external manipulation (e.g. through some sort of *API*).

Each of these steps in the pipeline will be discussed in more detail in the following sections.

A. Data Mapping

As previously mentioned, different municipalities might have different data models. On the other hand, it is also possible that the information contained in their Geographic Information Systems is incomplete or in a format that makes the Data Mapping process complex. There may also exist the need to relate several pieces of information or geometries in order to obtain new sets of data that are richer and more

complete. For this reason, there is a need for a flexible method to work on different scenarios.

To solve this problem, we introduce the concept of Mapping Module. Mapping Modules are capable of accessing GIS data sources and convert the stored data relative to a single type of urban entities into a format compatible with the Unified Model and store the new information in such model. They are also responsible for providing the transformed data to other Modules that require such data to function properly. This kind of dependency requires the Modules to be executed in a specific sequential order, where one Module can only depend on information provided by previously executed Modules. Which Modules are executed and in which order is specified, externally, by the user as it is specific to each situation.

As mentioned, one Module is responsible for mapping one and only one type of urban elements. However, these elements aren't always stored in the same way, leading to the need of different Modules to process the same kind of entity. The set of Mapping Modules that execute the mapping of the same type of entities is called a Family of Modules. Each Family provides methods that allow other Modules to query the processed data. There are several Families in the systems, corresponding directly to the types of entities in the Unified Model, specified in [8]. There are, for example, Families dedicated to mapping of Buildings, Roads, Vegetation and other Urban furniture.

Consider, for example, the data relative to the city terrain. One municipality might store such information as a cloud of points with height data associated, while another might store contour lines. There is clearly a need for two different algorithms to process each format and, as such, two Modules must be implemented. Since they map the same kind of data, they belong to the same Family of Modules, in this case the Terrain Family. After processing the respective data, both Modules can be queried, in the same fashion, for the height of a specific point in the terrain.

The Family of Modules concept creates a logic separation between different Modules, since one Module does not need to know about the specific implementation of the possible Modules on which it depends. All it needs to know is what Family it belongs to and then, call the respective methods. The system maintains a registry of Modules allowing the search for Modules based on their Family. Since it would make little sense mapping the same kind of elements more than once, the system only allows, at most, one Module of each Family to be running.

The execution of a Module starts with the retrieval of all the information from the data source. Afterwards, it processes this information, communicating with other Modules if necessary. Finally it stores the created elements in the Unified Model. Although it was possible to process the data while reading from the source, it may be desirable to treat the information as a whole and infer relations between all the urban elements. The processing phase works on both the geometric data (e.g., building lots) and the semantic data, and creates new entities with an axiom, which will be fed to the procedural modelling on a following step, and a set of attributes. However, in simple cases where a complex algorithmic component is not needed, the Mapping Module can be externally configured to create new attributes based on one of two actions: Copy and Set. The former, simply copies some data from the original data source to a new attribute, while the latter sets a new attribute with a specified value.

Sometimes, municipalities may store data that may prove to be uninteresting, such as information about buildings that are not yet built. If this data was allowed to be mapped into the Unified Model, it could lead to unexpected results like intersecting buildings or other incorrect urban elements. As such, Mapping Modules can be configured with filters to remove those elements from the process.

The Mapping Modules can also be configured through an *XML* file, allowing an arbitrary number of parameters to be passed to the specific algorithms and filters. This way, even if there is a need to develop new Modules for different cases, if the algorithms are general enough and can be configured through the *XML* file, in the long term we expect to gather a collection of Modules for the most common scenarios.

By the end of the Data Mapping stage, the Unified Model contains all the new urban elements needed to correctly model the urban environment. Each of these elements will be in a specific LOM, based on the amount of information that could be extracted from the GIS data source.

B. Data Conversion and Amplification

To facilitate the procedural modelling processes, these expect to be fed with the same amount of information for a given type of urban element. In other words, they expect the same LOM for all elements of a type of elements. However, as was seen before, these may be stored in the Unified Model in with an arbitrary Level of Mapping. The need for a mechanism to convert between LOMs is now apparent and thus the concept of LOM Converter was created.

LOM Converters are responsible for converting data between different levels of information, guaranteeing that the entities affected are output with a specific Level of Mapping. This is, they ensure that the entities will have at least the minimum semantic information required by the given LOM. Figure 2 shows a possible building modelled in LOM 1 and the same building after the LOM Converter was applied. LOM 1 only contains information regarding the building's volume while LOM4 contains information that is much more complete.

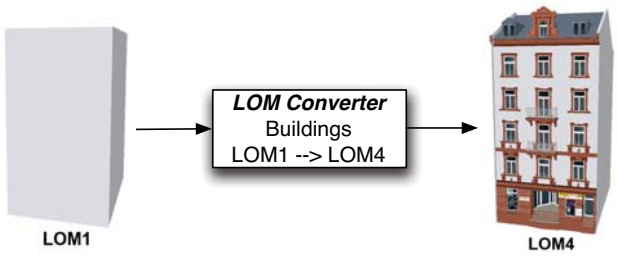


Fig. 2. Building LOM 1 to LOM 4 Converter

Given that the definition of Level of Mapping [8] does not impose a maximum level of information for a given LOM, the conversion from higher LOMs to lower ones does not make sense. However converting from lower LOMs to higher ones

requires the amplification of the existing information. This can be achieved through the introduction of randomness in this process, or with the use of heuristics.

Imagine the case where the production rules for buildings need information regarding roof types but there is no such information stored in the Unified Model. We can associate a random roof type to every building. However, this may lead to small houses with flat roofs or sky scrapers with gabble roofs. Although it is possible, it may seem unfamiliar. A better approach would be to define an heuristic and take into account the building height.

On the other hand, it might be necessary to create new production rules (e.g., because there is a need for a specific architecture) which may expect to be fed with more information. It is also possible that some attributes don't belong to any Level of Mapping. To accommodate for these scenarios, this stage in the pipeline can be configured through a file specifying how new attributes can be created. This file allows to specify random values for the attributes, copy them from those present in the Unified Model and relate the existing information to infer new data through the use of Data Amplification Operators which resemble those of a programming language, including relational operators, *if-then-else* statements and a domain specific operator which allows to specify a building's facade. This last one works by allowing the user to indicate the number of existing openings (windows and doors) in every floor and side of the building.

This way, the user can specify new data amplification mechanisms for specific cases. However, this file cannot alter the attributes created by the LOM Converters. In the same fashion, this file can also specify the file containing the production rules to model an element. This is particularly important when modelling monuments for instance.

This stage of the pipeline also takes the responsibility of manipulating CityEngine, turning it into a tool of the pipeline thus reducing human interaction. In other words, this step will control what operations CityEngine will execute with the information that it is being fed. These operations are, for example, to create a shape or a street segment, generate geometry or align the terrain to the shapes. These were called CityEngine Manipulation Operators. However, different cases may need different operations to take place in order to correctly generate the urban environment. As such, these operations and their sequence of execution can be configured externally using a file.

C. Modelling Processes

The last stage in the proposed pipeline is the execution of the procedural techniques which will generate the three dimensional models representing the urban space. These processes are fed with information from the previous steps. Normally this information consists of a geometric axiom (the shape) and a set of semantic attributes, but can also be a terrain or an attribute map. An example of the latter is the municipality's land use map which can control the types of buildings.

To create the three dimensional models representing the space, the procedural techniques are controlled by a set of CGA rule files, one for each type of urban elements. These, as mentioned before expect a set of attributes to be passed from the earlier stages of the pipeline. However, fallback values are commonly used.

Several times, it is desirable to control the Level of Detail of the produced geometries, either for efficiency reasons or to speed up the modelling process. As such, the production rules provide an attribute to control the Level of Detail. Also, several levels were defined for each of the types of urban elements.

As seen before, it is possible to define the file responsible for modelling an element. However, this can lead to CGA code being duplicated among files. Imagine, for example, that a city's characteristic building is being modelled and, therefore, needs a different rule file. Even though it may be quite different from other buildings, it may contain similar elements such as windows, doors or arcs, given they share the same architectural style. These elements could have the same code but are duplicated in both files. To prevent this, a library for each type of element capable of being reused was created. This libraries are simply rule files which have a rule that takes at least two parameters: the style and the Level of Detail. The former can be the architectural style in the case of building elements, or can be a type of urban furniture for example. The latter indicates the Level of Detail specified in the current element.

IV. Architecture

As mentioned before, the proposed pipeline can be implemented in a distributed architecture, where each of the discussed steps can take place in a different computer. Figure 3 illustrates this architecture.

Here, the information flows through all the components from the municipality's servers where GIS data is stored, to CityEngine where the transformed data is converted into three dimensional models reenacting the city's urban space. The Data Mapping stage is conducted by the application named *MappingEngine*. This is responsible for the sequential execution of the Mapping Modules which can connect to *WFS* and *PostGis* databases. Then, they transform the original data into an Unified Model compatible format. Afterwards, the transformed data is stored in the Unified Model.

To configure this step, the application relies on a set of *XML* files. The file *Modules.xml* contains information about which Modules to use and in which order should they be executed. It also specifies an extra configuration file for each Mapping Module. This is represented by the file *Module X.xml*. These files contain the connection parameters the Module needs to access the original data source, an arbitrary number of additional parameters to configure the specific algorithm, filters to remove unwanted data and information about how to map simple data. The file *DBConf.xml* contains connection parameters to access the Unified Model database.

On a different machine, the application *CEManipulator* is responsible for the Data Conversion and Amplification step. This application establishes a bridge between the Unified Model, CityEngine and the procedural modeling processes. Here is where LOM Converters are applied to every element retrieved from the Unified Model, effectively amplifying the existing data. It is also possible to generate more attributes,

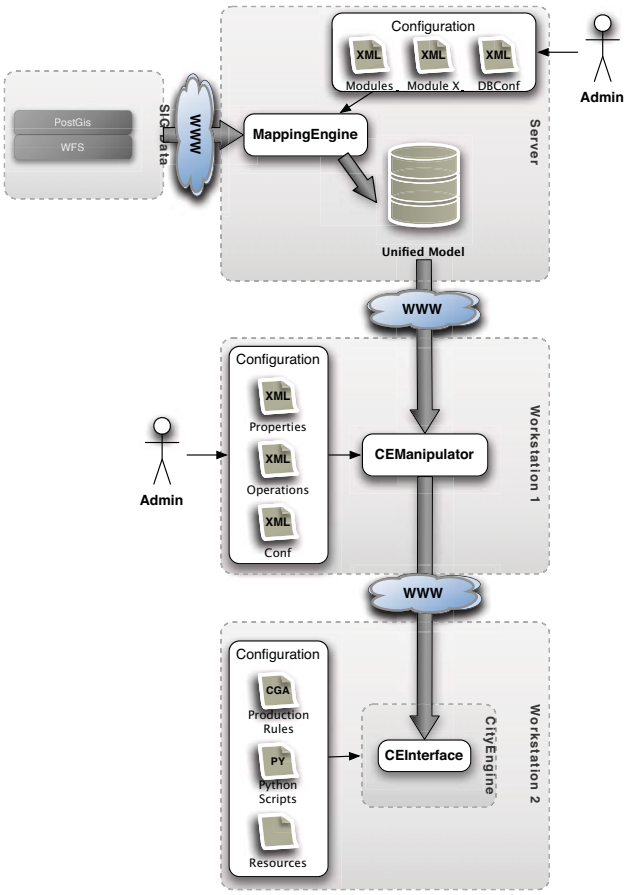


Fig. 3. Pipeline Architecture

other than those specified in the different Levels of Mapping, by applying the Data Amplification Operators. These are stored in the file *Properties.xml*. To control the procedural modeling processes that take place in CityEngine, this application counts with CityEngine Manipulation Operators defined in the file *Operations.xml*. The file *Conf.xml* contains connection parameters to access both the Unified Model and the plugin *CEInterface*, described next.

To allow the external manipulation of CityEngine, a plugin was developed using the *Python* scripting interface. This plugin implements a server that receives the operations from *CEManipulator* and, executes the respective actions. Because it is a server, it allows the CityEngine to be a tool for the pipeline and thus, reducing the need for human interaction. This way, it is possible to remotely execute the procedural modeling processes and retrieve the generated models without physical access to the computer running CityEngine.

V. Results

In this section we present the results obtained using the proposed pipeline to model the urban space of the municipality of Santa Maria da Feira, in Portugal. We also present screenshot images displaying the models generated being used in the Unity 3D Game Engine in an interactive application.

A. Efficiency

It is possible that urban environments may have thousands of elements that one may wish to model. For this reason, if care is not taken, the pipeline's execution may take a lot of time to produce results. As such, it is important to have some kind of metric regarding the time spent by the pipeline to generate urban environments, measuring the individual times for each step.

Table I we shows the times relative to the Data Mapping stage. This table shows information regarding the mapping of buildings, roads and the whole city. Here, $N\neg\int$ *Read* indicates the number of elements read from the original data source, $N\neg\int$ *Written* indicates the number of elements written to the Unified Model and *Time* represents the time spent on this process. We can see that the number of read items and the written items isn't the same for each type of entities. This is caused by the automatic removal (using filters) of non interesting data.

TABLE I. Times relative to the Data Mapping.

Type	$N\neg\int$ Read	$N\neg\int$ Written	Time
Buildings	1822	1640	116.8s
Roads	282	228	178.58s
Whole City	14631	9899	359.4s

These values depend a lot on the format of the original data and the algorithms involved in processing such data. From this point on, the rest of the pipeline is not significantly affected by this format.

In the following Table, the times for the Data Conversion and Amplification is shown. As in the previous table, this one presents information for buildings, roads and the entire city. $N\neg\int$ *Read* represents the number of elements read from the Unified Model, *Imported* indicates the number of CityEngine elements (Shapes and Graph Segments) created, and *Time* represents the time spent on this process for each type of entity. The number of roads imported differs from the number of roads retrieved form the Unified Model because of the different representations of this type of data. The Unified Model treats one roads as one *Polyline* element, while CityEngine treats them as a set of graph edges as nodes.

TABLE II. Times for Data Conversion and Amplification.

Type	$N\neg\int$ Read	Imported	Time
Buildings	1640	1640	24.2s
Roads	228	4972	395s
Whole City	9035	10941	1221s

Table III contains information about the time spent by CityEngine modeling buildings, roads and the whole city. $N\neg\int$ *Elements* indicates the number of elements to be modeled by CityEngine, $N\neg\int$ *Polygons* indicates the total number of polygons generated and *Time* indicates the time spent in the procedural modeling.

TABLE III. Times for procedural modeling.

Type	$N\neg\int$ Elements	$N\neg\int$ Polygons	Time
Buildings	1640	182574	76.1s
Roads	4972	5352	1.9s
Whole City	10941	4847969	165.4s

Summing the times spent on all three steps for the entire city is possible to conclude that the whole process takes around thirty minutes to complete for a city of 14631 elements. This is an average of 0.16 seconds per city element.

B. Procedural Modeling

The primary goal of this pipeline is to procedurally generate three dimensional models that resemble an existing urban space. As such, it is necessary to analyze these processes regarding the level of detail, visual fidelity and geospatial contextualization.

To achieve a certain level of visual quality, it is important that procedural generation tools are capable of creating elements with an adequate level of detail. In Figure 4 we present a building modeled with CityEngine. The model in the Figure, represents a house with detailed windows and doors. Also, CityEngine is capable of incorporating previously modeled elements into its models allowing the combination of procedural techniques with traditional ones, increasing the level of detail provided by this tool.

Fig. 4. House modeled with CityEngine

Visual fidelity is a very important factor in the representation of existing urban environments. It also plays a critical role by allowing users to recognize the spaces. The correct positioning and spatial relation between elements and their height and shape contributes a lot to establish some relation with the real spaces. In the case of buildings, the production rules created allow some control over the aspect of their facades, by allowing the number of openings (doors and windows) to be specified for each floor in every facade, by changing an attribute. Figure 5 shows the same building with different openings by floor and face. In this image it is possible to see that, on the left, all facades and floors have three openings, while on the right the same house has two windows on the top floor of the front side, and one door and two windows on the first floor. Moreover, the right one looks more realistic and has a better visual appearance. By allowing such control, it is possible to model certain buildings to look more like their real counterparts.

Geospatial awareness allows to create relations between different elements in an urban space, allowing to infer new

Fig. 5. Same house with different openings.

properties from the surrounding elements. To this date, however, CityEngine only provides means of verifying if an element intersects another one and to check if the side of a building is oriented towards a street. In the latter case, this spatial awareness is only automatically possible when it is CityEngine itself creating the building lots. In other cases, however, it is possible to use *Python* scripts to define which are the sides nearest to a street. For more complicated cases, where a more complex algorithmic component is needed, it is preferable to use Mapping Modules to solve geospatial awareness problems.

C. Interactive Application

As a proof of concept, we have developed an interactive application using Unity 3D Game Engine, allowing the user to freely navigate the generated urban virtual environment. This Game Engine was chosen for its capability to operate on large terrains. Also, it is quite easy to integrate the generated models and terrain with Unity. However, some terrain areas might require minor modifications, as it is possible that it is not leveled with the remaining models. It's fairly easy to detect these situations as the models may appear partially buried in the ground or floating above it. Figure 6 presents some screenshots obtained in Unity 3D.

VI. Conclusions and Future Work

In this paper we propose a three stage pipeline for the procedural modeling of real urban environments, based on information stored in Geographic Information Systems. Such pipeline aims to create a bridge between GIS maintained by municipalities and procedural modeling techniques, harvesting the power of both technologies. On one hand, Geographic Information Systems provide information that can be fed into procedural modeling, allowing to generate virtual urban environments that are more similar to the real ones. On the other, procedural techniques can accelerate the modeling of such spaces, since they require less human interaction while also reducing the costs associated with such task.

Distinct data models are mapped into an Unified Urban Model according to a discrete number Levels of Mapping. Therefore it is possible to generate virtual environments of distinct municipalities with the same modeling processes, just by adjusting the mapping of semantic information at the first stage of the pipeline. Even if we have data from different data sources classified in distinct LOM, we can amplify these data into an unique resulting LOM on the second stage of this pipeline. Therefore reusability is promoted for distinct GIS data models or the integration of distinct data sources into an unified virtual urban environment.

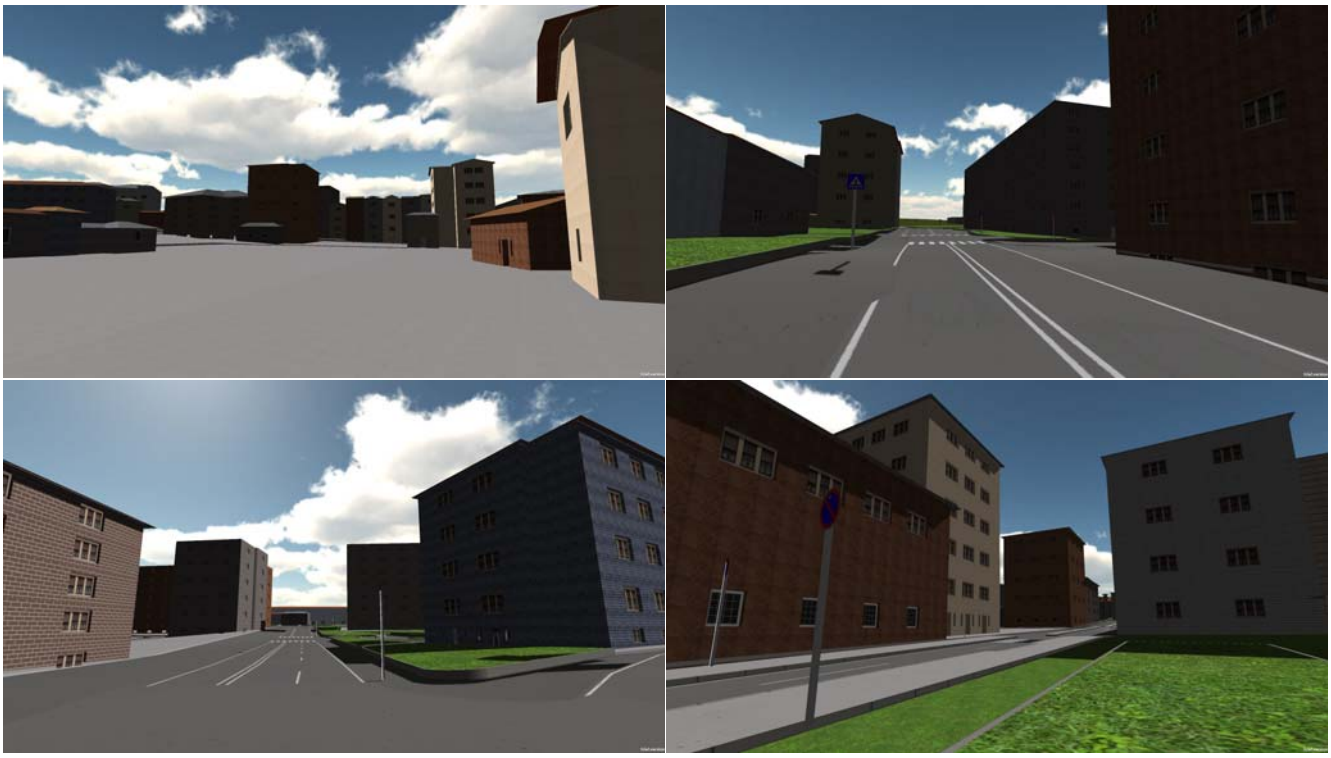


Fig. 6. Interactive application in Unity 3D Game Engine.

The Mapping Modules were developed specifically for this case study and, as such, there is a need for more generic and configurable Modules. In the same way, the development of LOM Converters for all entities and mapping levels is also required. Due to the large amount of configuration the pipeline needs, the development of a graphical user interface for that purpose would be interesting. Further research on metadata will provide the automation of the mapping process.

REFERENCES

[1] Procedural., "Esri cityengine — 3d modeling software for urban environments," http://www.esri.com/software/cityengine, 2012.

[2] Y. I. H. Parish and P. Müller, "Procedural modeling of cities," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '01. New York, NY, USA: ACM, 2001, pp. 301–308. [Online]. Available: http://doi.acm.org/10.1145/383259.383292

[3] A. Lindenmayer, "Mathematical models for cellular interaction in development: Parts i and ii." *Journal of Theoretical Biology*, vol. 18, 1968.

[4] R. Měch and P. Prusinkiewicz, "Visual models of plants interacting with their environment," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '96. New York, NY, USA: ACM, 1996, pp. 397–410. [Online]. Available: http://doi.acm.org/10.1145/237170.237279

[5] P. Wonka, M. Wimmer, F. Sillion, and W. Ribarsky, "Instant architecture," in *ACM SIGGRAPH 2003 Papers*, ser. SIGGRAPH '03. New York, NY, USA: ACM, 2003, pp. 669–677. [Online]. Available: http://doi.acm.org/10.1145/1201775.882324

[6] P. Müller, P. Wonka, S. Haegler, A. Ulmer, and L. Van Gool, "Procedural modeling of buildings," *ACM Trans. Graph.*, vol. 25, pp. 614–623, July 2006. [Online]. Available: http://doi.acm.org/10.1145/1141911.1141931

[7] A. Coelho, M. Bessa, A. A. Sousa, and F. N. Ferreira, "Expeditious modelling of virtual urban environments with geospatial l-systems," *Computer Graphics Forum*, vol. 26, no. 4, pp. 769–782, 2007. [Online]. Available: http://www3.interscience.wiley.com/journal/118494430/abstract

[8] T. Martins, "Ontologia urbana para ambiente virtual colaborativo no contexto do planeamento e gestão municipal," Master's Thesis, Faculdade de Engenharia da Universidade do Porto, July 2011.

An Approach to Edge Detection in Images of Skin Lesions by Chan-Vese Model

Roberta B. Oliveira, João Manuel R. S. Tavares
Instituto de Engenharia Mecânica e Gestão Industrial
Faculdade de Engenharia da Universidade do Porto, FEUP
Porto, Portugal
robyoliveira1@gmail.com, tavares@fe.up.pt

Norian Marranghello, Aledir S. Pereira
Departamento de Ciências de Computação e Estatística
Instituto de Biociências, Letras e Ciências Exatas, UNESP
São José do Rio Preto, Brazil
norian@ibilce.unesp.br, aledir@ibilce.unesp.br

***Abstract*—Nowadays there is a great interest in the application of computational systems for the analysis of skin lesions. These systems allow the dermatologist to prevent the development of malignant lesions. The development of the systems has occurred due to the increase of skin cancer cases. In the characterization of skin lesions it is necessary to segment the images accurately. Thus the features and edges information of the lesion can be extracted and used by a classifier or by a dermatologist for a better classification. When images are acquired in a non-systematic and non-controlled way there may be a segmentation problem. In this case the skin lesion of images can have different sizes and various type of noises, such as the hair. These factors can affect the detection of the lesion edges and complicate its characterization. One solution would be to apply a smoothing filter to reduce noise before the segmentation step. Segmentation techniques adapted to each type of image can be used to solve the problem of diversified images, such as images with different sizes lesions, reflexions and light intensities. In this paper is proposed a computational method to assist the dermatologists in the diagnosis of skin lesions by digital images. It was used the anisotropic diffusion technique for the preprocessing of the images in order to remove the noises. The Chan-Vese model was used to segment the lesions. The next step consists of the application of morphological filters to eliminate outside and inside noises from the object, that remained in the segmented images, and also to smooth their edges. This approach allowed to minimize noise problems and edge detection to different cases of skin lesions images, such as melanoma, melanocytic nevi and seborrheic keratosis. The segmentation achieved 94.36% of accuracy for the three types of skin lesions.**

Keywords—Skin Lesions; Anisotropic Diffusion Filter; Chan-Vese Model; Morphological Filters .

I. INTRODUCTION

The appearance of skin cancer can happen for several reasons, one of them is excessive exposure to the sun, a preventive action against this situation is needed. However great attention should be given to nevi (moles), that are benign lesions, being 50% of melanoma cases derived from moles [22]. It is also important to know that melanoma can resemble a mole when in its initial state. Another benign lesion that is important to analyze is seborrheic keratosis. In some cases, the diagnosis of this lesion is confused with melanoma, so the differentiation of these two types of lesion is important.

Melanoma is the most aggressive skin cancer, because it has a high mortality rate. Nevertheless when it is diagnosed early and treated properly, the cure of patients with this type of cancer can reach 69% of world average [6]. The increase in cancer cases has motivated the research and development of computational methods to assist dermatologists in the diagnosis of skin lesions. The goal is to analyze the benign lesions to prevent their development, or diagnose malignant lesions at their early stage, so they can be treated early and increase the chances of a cure.

A problem in the skin lesions segmentation step is influence of noises, such as hairs, that could impede the lesions segmentation. Another problem is the heterogeneity of database images that can decrease the efficiency of the used segmentation method. The smoothing technique is important to minimize the problem of noises in the images. In the case of heterogeneity of the database images, the use of a segmentation technique is very important to enable the efficiency of edges detection. It allows the adaptation to the problem of each individual image, such as images with different sizes lesions, reflexions and light intensities. The approach proposed in this paper to solve the presented segmentation problems is based on anisotropic diffusion technique [3] which reduce the effect of noises on the images and active contour model without edges (Chan-Vese model) [7] to identify the diseased area. This model is derived from a compilation of two techniques: the Mumford-Shah region growing technique, used to segment the images; and also the Level Set Active Contour Model, which allows topological change of the curves, used to edge detection. The use of these techniques is subject of several papers for edge detection of skin lesions [1, 2, 4, 21, 24, 25].

The objective of this paper is focused on the development of a method for edge detection of skin lesions, such as, melanoma, melanocytic nevus and seborrheic keratosis, from photographic images to assist dermatologists in their diagnoses. It is expected that this approach allows an accurate detection for most lesions and thereby the edge information can be made available to the dermatologist or use by a classifier.

The second section discusses some papers related to the theme. The third section explores the proposed method. The fourth section is a discussion of the results obtained. In the fifth section presents a conclusion of the approach developed and future work.

II. LITERATURE REVIEW

Considering the importance of early diagnosis of skin lesions there are several papers that propose automatic methods to assist dermatologists in their diagnoses. These papers show digital image processing techniques to segment different skin lesion types.

Beuren and collaborators [5] propose a morphological approach to melanoma image segmentation. The filtering of images is started by appling a morphological opening process, to eliminate the hairs and others noises. Subsequently it is applied a global thresholding, where filtered images by color are binarized. The binary opening filter is used to fill gaps in the segmenting region and remove the external noises. This paper evaluates segmentation of 200 benign and malignant lesions images. They obtained 95.26% of accuracy for benign lesions and 92.62% for malignant lesions.

A method for edge detection in dermoscopic images of melanocytic and non-melanocytic lesions is proposed by Norton et al [17]. In this paper was carried out two segmentation: general lesion segmentation and the bright region segmentation. The general lesion segmentation is was performed in three steps: I) segmentation of tumor area, II) correction of non-uniform lighting and III) noise reduction. For general lesion segmentation they used thresholding and to bright region segmentation, correction of lighting and noise reduction was applied morphological operators. The evaluation of this method was based on the accuracy with 84.5% for 107 images of non-melanocytic lesions and 93.9% for 319 images of melanocytic lesions.

Cudek et. al. [8] provide a method for identifying skin lesions from digital images using the ABCD rule. They used 53 images of skin lesions, which were rotated 90, 180 and 270 degree, forming a database of 212 images, which include benign nevi, blue nevi, suspicious nevi (dysplastic) and melanoma, which were transformed into gray levels . The technique of histogram equalization is used to enhance the contrast of images, and the median filter was applied to the to decrease the noise. A modification of the Otsu [19] threshold was proposed for segmentation of the skin lesions, because it was examined that this method may, in some cases, ignore regions that compose the lesion. Thus, to solve this problem it was proposed SD (Small Difference), to search pixels in the neighborhood that can be classified as regions of the lesion. The proposed segmentation obtained 92% of correct detection. In 5% of the images it was necessary to manually enter the threshold and in 3% of cases was incorrect recognition was observed.

III. METHODOLOGY

This section presents the developed approach for edges of skin lesions detection, in order to assist dermatologists in their diagnosis. The structure of the method can be seen in Fig. 1.

The method has the following processing steps: image acquisition, preprocessing, segmentation, post-processing images. Firstly images are acquired to compose the image database. The second step of the proposed method consists in the smoothing of the images using anisotropic diffusion filter to remove noise, which interfere with the following technique, the image segmentation. Subsequently the segmentation is performed using the active contour model without edges, Chan-Vese model, to detect the lesion. Finally, morphological filters are applied in the segmented images to soften the edge and remove noise resulting from the process of segmentation.

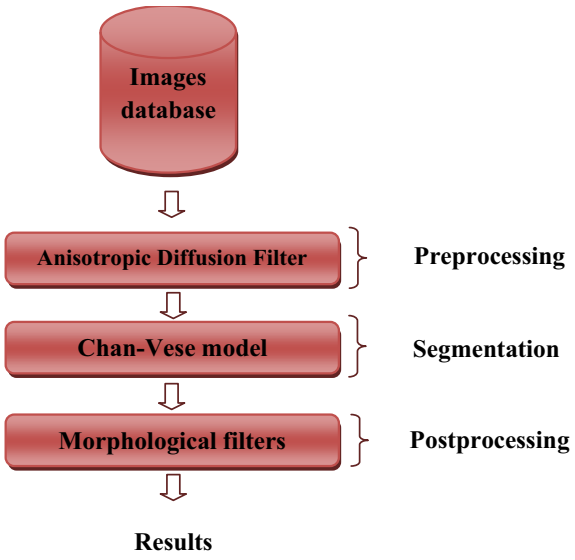
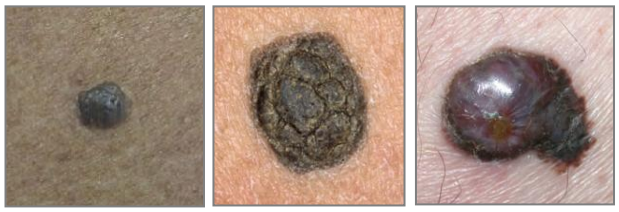


Fig. 1. Developed approach structure.

A. Image database

The first step of the developed method was image acquisition to form an image database of skin lesions, used for testing. The database of this paper consists of images of the following bases: Loyola University Chicago [15] YSP Dermatology Image Database [10], DermAtlas [9] DermIS [11] Saúde Total [22], Skin Cancer Guide [23] and Dermnet - Skin Disease Atlas [12, 13].

The database used in this paper consists of 408 images, 62 melanocytic nevi images, 86 seborrheic keratosis images and 260 melanoma images. The high amount of melanoma images arises due to the raised concern with this type of lesion, which is carcinogenic and has a high mortality rate. The dimensions of the images of the database is 200×200 pixels. Examples of each of these types of skin lesion are shown in Fig 2.



(a) (b) (c)

Fig. 2. Images database: (a) melanocytic nevi, (b) seborrheic keratosis and (c) Melanoma.

B. Preprocessing

In this step it was applied a smoothing technique in the database images, in order to soften the effects of noise, which may hinder the segmentation result. The nonlinear filter used to

smooth was the anisotropic diffusion, as proposed by Barcelos, Bonaventure and Silva [3]. This filter was chosen because of its good result in images smoothing and preservation the edges of the skin lesions [4]. Since one of the characteristics analyzed for the diagnosis of skin lesions is the irregularity of the edges, this filter proves to be an efficient technique to smooth the images.

The segmentation method used in this paper is applied to images in gray levels. Thus, the colored original images are converted to gray levels. Thus, the processing time is decreased since it is not necessary to smooth the three RGB components separately. The images are smoothed by anisotropic diffusion method.

The implementation of this filter is based on (1):

$$u_t = g\,|\nabla \mathrm{u}|\; div\left(\frac{\nabla u}{|\nabla \mathrm{u}|}\right) - \lambda\,(1-g)(u-I) \tag{1}$$

$$g = \frac{1}{1+k|\nabla(G_\sigma * u)|^2} \tag{2}$$

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3}$$

where $u(x,y,0) = I(x,y)$; u is the smoothed image; I is the original image; $t > 0$ is the smoothing scale; div is the divergence operator; ∇u is the gradient value of u; λ is a parameter which assist in the speed of diffusion.

The term g presented in (1) is defined in (2). The term g is used for edge detection, where k is a parameter, the term G_σ is the Gaussian function that it is presented in (3) and σ is the standard deviation of G_σ. Considering a neighborhood of a point x, when the gradient ∇ has a low average, ie, there are few points (noise) in the image, the x is considered an interior point (homogeneous region) resulting in $g \sim 1$. But, if the gradient ∇ has a high average, having several points, x will be a point of a contour, $g \sim 0$.

C. Segmentation

Skin lesions segmentation is used to separate the diseased area of the healthy region, which promotes the detection of the lesion. The technique used for segmentation of the images in this paper was the active contour model without edges proposed by Chan and Vese [7] Thus, this method was applied for images in gray levels. This method is performed by energy minimization of the curve over the image. The segmentation of this model is based on region and uses the concepts of Mumford-Shah techniques [16], commonly used in image segmentation tasks. The segmentation of the model is also based in the concepts of level set active contour model [18], which allows topological changes applied on the input images. The Chan-Vese model is applied as follows:

$$F(c_1,c_2,\phi) = \mu \int_\Omega \delta\big(\phi(x,y)\big)|\nabla\phi(x,y)|dxdy \tag{4}$$

$$+\nu \int_\Omega H\big(\phi(x,y)\big)dxdy \tag{5}$$

$$+ \lambda_1 \int_\Omega |u_0(x,y) - c_1|^2 H\big(\phi(x,y)\big)dxdy \tag{6}$$

$$+ \lambda_2 \int_\Omega |u_0(x,y) - c_2|^2 (1 - H(\phi(x,y)))\,dxdy \tag{7}$$

where having as fixed parameters μ, $\nu \geq 0$ and λ_1 and $\lambda_2 > 0$ and the term u_0 is smoothed image. The constants c_1 e c_2 are the average image u_0 inside the curve C and the average outside of the curve C, respectively. that assist each term in its results, are expressed by:

$$c_1(\phi) = \frac{\int_\Omega u_0(x,y)H(\phi(x,y))dxdy}{\int_\Omega H(\phi(x,y))dxdy} \tag{8}$$

$$c_2(\phi) = \frac{\int_\Omega u_0(x,y)(1-H(\phi(x,y)))dxdy}{\int_\Omega (1-H(\phi(x,y)))dxdy} \tag{9}$$

where H e δ are the Dirac and Heaviside function, respectively, to obtain the level set energy function $F(c_1,c_2,\phi)$ [18].

There are several advantages of this method, which allows its use to provide good results. It allows the detection of different objects with different intensities and also with blurred boundaries; it allows the curve topological change; it allows object detection where the contour has no gradient, due to the stop criteria of the curve evolution until the desired boundary not depending on the image gradient; and it allows objects detection in noisy images [7].

The first stage of segmentation using the Chan-Vese model is the definition of a curve over the image, which will be minimized to the border of the object. This model has features like the ability to place the curve anywhere on the image, representing it in various forms and different sizes.

In this paper the initial shape of the curve was a square and it was positioned close to the image center, so a small number of iterations are performed for the curve involve lesion which reduces the processing time. Initially, the curve was defined as 140×140 pixels, whereas most images, approximately 64% are composed of larger lesions, in other words, they represent a large number of pixels. This tends to occur when the distance at the acquisition time is very close to the lesion. However, we observed that closer the curve better the results and lesion detection. Moreover smaller is the chance to find false edges, such as those caused by reflections and shadows. Therefore, considering the significant amount of small lesions, it was necessary to define a smaller curve for small lesions and a bigger curve for large lesions. Two curves were established with different sizes: 40×40 pixels and 140×140 pixel to put them over small and large lesions, respectively. The definition of the size of the curve should be informed by the user as needed.

With the intention of automating the definition of the curve for the segmentation process by Chan-Vese model, a method was analyzed to establish the curve according to the size of the lesion.

For the definition of the threshold that is used to establish the curve, we considered only images composed by small lesions, in which the use of the curve with size 40×40 pixels obtained better results by applying the Chan-Vese model. Firstly, it is made to count only those pixels that are part of the lesion from the segmented image, ie, the black pixels. After this step, the average and the standard deviation of the pixels pertaining to all images of small lesions is calculated. The sum of the average with standard deviation define the threshold (L). The threshold enables to differentiate the small lesions in the large segmented images. Considering the dimensions of the images of the database 200×200 pixels, the images are composed of 40,000 pixels. The result of the threshold was 6345 pixels, which represents the amount of pixels comprising the images with small lesions.

Considering the definition of the threshold by analysis of manual segmentation using the Chan-Vese model, it was possible to automate the definition of the curve for images segmentation. So, the thresholding technique is applied, by the method Otsu [19]. For each binarized image is counted the number of pixels (TP) corresponding to the lesion. For the definition of the curve it is analyzed the threshold. If the total number of pixels is less than or equal to the threshold ($TP \leq L$), the curve defined over the lesion is small, 40×40 pixels. If the value is greater than the threshold ($TP \leq L$), the curve will be large, 140×140 pixels, for the application of the Chan-Vese model.

D. Postprocessing

After performing the segmentation step by Chan-Vese model, the morphological filters was applied on the binarized images. The use of filters eliminates inside and outside noise from the segmented regions. These noises, such reflexes, can cause the definition of the false border by the segmentation method.

The filters used in this paper were "opening filters" and "closing filters" [14]. The structuring element has the shape of ellipse, the two radiuses are equal to four. The filter application allowed to smooth the edge and eliminate internal or external noises of the lesion.

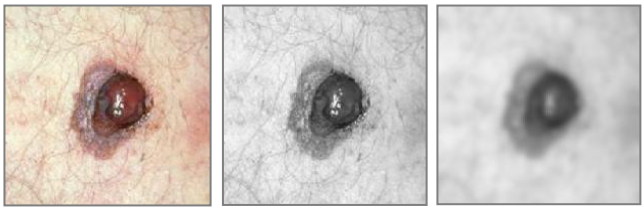
IV. RESULTS AND DISCUSSIONS

The goal of the tests is to evaluate the approach of the method for edge detection of skin lesions, which will assist the dermatologist in their diagnosis. All database images were used to evaluate results. In this stage were applied the anisotropic diffusion technique to smooth images and Chan-Vese model for the segmentation of skin lesions. The contours established by the method were evaluated visually by dermatologist Dr. Ricardo Rossetti Baccaro.

A. Anisotropic diffusion

The implementation of this filter was based on the discretization of (1), which has the following parameters: Δt determines the size of the temporal evolution, wherein each iteration of the diffusion will be executed; and the parameter σ is the standard deviation of the Gaussian function G_σ; the parameter λ assist reinforce the edge; k assist the Gaussian function to define if the point is part of the edge or not. If it is the edge point, this point will be less smoothed. This smoothing will be in accordance with the number of iterations NI.

The result of applying this filter can be seen in Fig. 3. The parameters were determined by tests, considering the parameters already defined in the paper [4], with the following values $\Delta t = 0.1$, $\sigma = 1$, $\lambda = 1$, $k = 0.0008$ and $NI = 100$.



(a) (b) (c)

Fig. 3. Result of applying the anisotropic diffusion filter. In (a) original image, (b) gray levels image and (c) smoothed image.

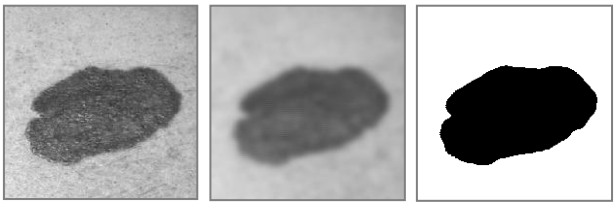
The original image (a) of Fig. 3 is converted to gray level image, which can be seen in the image (b). In (c) we have gray levels image, smoothed by anisotropic diffusion filter. We can see that the filter removed the presence of hairs in the images, yielding promising results. In the case of images having shadows or reflections areas, as in this case, the filter did not permit to eliminate influences thereof, although their presence was ameliorated.

B. Chan-Vese model

In the application of the Chan-Vese model the discretization of (4) is used for the evolution of the curve. In which the parameters were defined by tests, considering the parameters already defined in the paper [7]:

- $\mu = 0.2$, parameter that controls the length of the curve;
- $\nu = 0$, influence in the area inside the curve;
- λ_1 e $\lambda_2 = 1$, assists in locating the object inside and outside the curve respectively;
- $h = 1$, assists in detection of inner contour;
- $\Delta t = 0.1$, is the time of evolution of the curve.

500 iterations were applied to the evolution of the curve, in other words, the minimization of the curve occur until the number of defined iterations or when the curve is located on the object. In Fig. 4 is shown the result of applying the Chan-Vese model from smoothed image (b), that allowed the image to be binarized, as it can be seen in the image (c).



(a) (b) (c)

Fig. 4. Result of applying the Chan-Vese model. In (a) gray levels image, (b) smoothed image and (c) binarized image.

Some images resulting from the segmentation process present: holes inside of the lesion area and the outside noise, caused by reflections, shadows or some other noises that were not eliminated in the step of smoothing. These factors are addressed in the next step of post-processing of images.

C. *Morphological filters*

In (c) of Fig. 5 can be seen the effect of the application of morphological filters resulting from "opening" followed by "closing" the segmented image (b) by Chan-Vese model. We can see that the lesion holes of the image (b) were eliminated and it also softened the edge without losing its features.

(a) (b) (c)

Fig. 5. Application of morphological filters. In (a) smoothed image, (b) binarized image and postprocessed image.

D. *Edge detection*

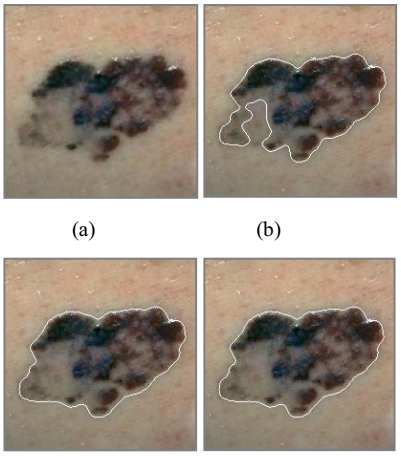
After performed post-processing step in the skin lesion images, its edge is defined. The edge represents the boundary and irregularities lesion, allowed the separation of the diseased region from the healthy region.

The edge is define by pixels where sudden changes occur in the intensity of binarized images. In this paper the lesions are represented by black color and the skin are represented by white color. Scanning the image pixel by pixel, when there is a change in color, the pixel at the same position in the original image receives white color to define the edge of lesion.

In order to obtain a result that best represents the edges of the lesion, tests were conducted to evaluate two techniques for skin lesions segmentation: the thresholding and the Chan-Vese model. Both techniques were applied to the smoothed images by anisotropic diffusion. Then it was defined the edges of the lesions from the postprocessed images by morphological filters.

The image segmentation by thresholding technique is applied by OTSU threshold [19]. The lower intensities of the threshold are defined by "0" to represent the lesion. Whereas, greater intensities of the threshold receive "1" to represent the skin. In the case of segmentation by Chan-Vese model, it was used the same parameters mentioned above for the evolution of the curve. The curve is established depending on the size of the lesion. If the lesion is small, the size of the curve is defined by 40×40 pixels, and if the lesion is big, the curve is defined by 140×140 pixels.The proposed method for automating the definition of the curve also was evaluated. This method utilizes the amount of pixels that composes the lesion returned by thresholding method to define the size of the curve based on a threshold. The threshold is established by mean and standard deviation of points belonging to the lesion, that were returned by application of the most appropriate Chan-Vese model to the large and small images.

In Fig. 6 is showed the results of the three segmentation methods in the original image (a). The segmentation result using thresholding technique is presented in the image (b), the result of applying the Chan-Vese model can be seen in the image (c) and in (d) the contour was obtained by using the automating method for definition of the curve by Chan-Vese model. We can observe in the image (c), where it was applied the Chan-Vese model, that the edge surrounded better the lesion than in the image (b) obtained by thresholding technique. Some regions of the lesion in the image (b), were not completely detected. In the case of image (d), the result was the same as the edge of the image (c), since the same size was determined to the curve by automated method, using Chan-Vese model.



(a) (b)

(c) (d)

Fig. 6. Results of segmentation methods. In (a) original image, (b) result using thresholding technique, (c) result using Chan-Vese model and (d) result using the automating method for definition of the curve by Chan-Vese model.

The results of the Chan-Vese model surrounded better the regions with intensities closer to skin color than the threshold technique. In some cases, the method for establishing an automatic curve does not detect correctly the lesions due to the wrong definition of the size of the curve. For example, when the curve should be smaller, 40×40, the amount of pixel that comprises the lesion obtained by thresholding method was greater than the threshold. For this reason, there are cases in which correctly detects the lesion, but also identified other regions, due to the size of the curve.

For very noisy images, Chan-Vese model also achieved good results with the influence of the anisotropic diffusion filter, unlike the results of applying thresholding technique with which most lesions were not detected (Fig.7).

The images that have some regions of the skin such as redness, shadow and reflection, can influence the detection of the edges of the lesion. In such cases the regions are regarded as belonging to the lesion by thresholding technique, unlike the Chan-Vese model which detected only the skin lesion.

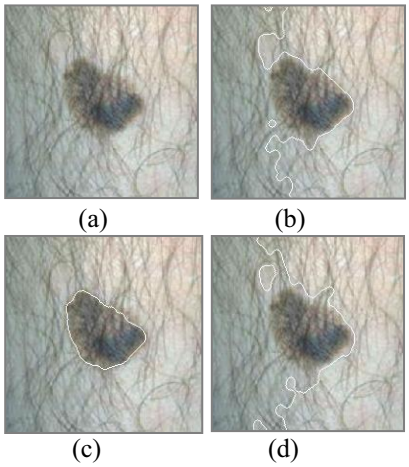


Fig. 7. Segmentation of noisy images. In (a) original image, (b) result using thresholding technique, (c) result using Chan-Vese model and (d) result using the automating method for definition of the curve by Chan-Vese model.

We evaluated the results based on all images of database: 62 melanocytic nevi, 86 seborrheic keratosis and 260 melanoma. The images were analyzed by an expert to identify if the skin lesion edge was detected or was not detected by segmentation techniques: thresholding, Chan-Vese model and also with the application of the automatic curve. The lesions were considered detected by dermatologist in the cases where the lesion have been fully engaged by the edge.

According to the results shown in Fig 8, the aplication of Chan-Vese model gave best results for the segmentation of skin lesions than the other techniques also analyzed. This model allowed to correctly detect 94.36% of the images. The thresholding technique had the least significant result, with 80.39% of the images correctly detecting the edge of the skin lesions. This result occurred due to the difficulties in the segmention of images with noise, such as shadows, even though using the anisotropic diffusion smoothing method,

The proposed method to define a automatic curve in the segmentation by Chan-Vese model, achieved a better result than the threshold technique. Approximately 14% of the images were not detected due to incorrect setting of the size of the curve, which is based on the thresholding technique. This problem usually happen when the lesions are considered small which requires a small curve for its detection. In these cases it is defined a large curvebecause the thresholding method does not correctly detect the lesion which determines a large area in the segmented images. This large area becomes greater than the threshold needed to establish a small curve in the image.

The incorrect definition of the curve, caused error in the segmentation of certain images, not allowing proper detection of the edges lesions, due to the lower number of iterations for this type of situation. For this case, the minimization of the curve require more iterations and could also detect other false objects in the image due its proximity to the curve.

The seborrheic keratosis lesions type present higher detection errors, due to similar colors between the lesions and the skin, which make the lesions dificult to detect. Another examples where the lesions are not detected, where the images where reflections and shadows were not eliminated in preprocessing.

The proposed method proved to be a promising technique for images with many hairs and intensities near of skin color. In some images were there found a few false edges, due to some stains or noises, when those were not completely eliminated by the smoothing filter.

The papers discussed previously [5, 8, 17], which apply the segmentation by thresholding techniques and the results of this study using the Chan-Vese model are shown in Table I, It is also presented the references of papers, years of their publications, methods of segmentation, types of lesions addressed and their accurancy, respectively.

TABLE I. RESULTS OF PAPERS FOR SEGMENTATIONS SKIN LESIONS.

Source	Year	Segmentation	Images	Accuracy
Proposed approach	2012	Chan-Vese model	Melanocytic nevus	96.77%
			Seborrheic keratosis	93.02%
			Melanoma	94.23%
[5]	2011	Thresholding	Benign lesions	95.26%
			Malignant lesions	92.62%
[17]	2010	Thresholding	Non-melanocytic lesions	84.5%
			Melanocytic lesions	93.9%
[8]	2010	Thresholding	Benign and malignant lesions	92%

In the first paper [5] the authors obtained 95.26% of accuracy for benign lesions and 92.62% of accuracy for malign lesions. For the second study [17], the authors segmented the melanocytic and non-melanocytic lesions, which received 84.5% and 93.9% of accuracy, respectively. In the third study [8], the authors evaluate the accuracy for benign and malignant lesions, and obtained 92% accuracy considering the two classes. In this paper, the proposed method achieved 96.77% of accuracy in segmentation melanocytic nevi, 93.03% of accuracy for the class of seborrheic keratosis and 94.23% of accuracy for the class of melanoma.

It was impossible to perform a comparison between the different studies because of several factors. The images database used in which study were different, the addressed skin lesions were not of the same type as well as the application of the segmentation thecniques which was shown to be also different.

The results presented in this study show that the developed method achieved promising results in segmentation through the Chan-Vese model for skin lesions of melanocytic nevus, seborrheic keratosis and melanoma types.

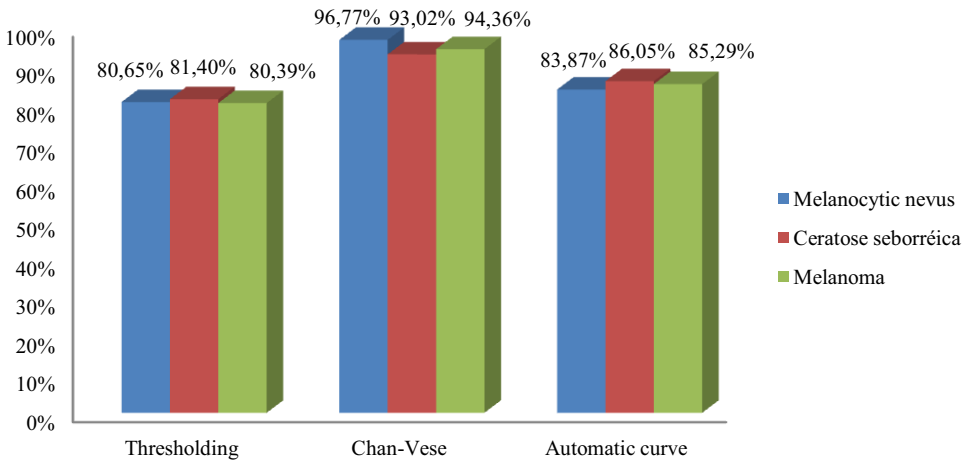


Fig. 8. Comparison of segmentation methods.

V. CONCLUSIONS AND FUTURE WORK

This paper presented a method for edge detection of skin lesions, using the Chan-Vese model to provide information about the edges of the skin lesions, in order to assist the dermatologist in his diagnosis. To minimize the presence of noise on the images was used the anisotropic diffusion non-linear filter. The active contour model without edges was applied in order to establish a better edge to the lesion.

The application of anisotropic diffusion filter was held to smooth images without losing the rough edge. The noises were partially eliminated, except in cases with presence of reflections and shadows. This filter was essential to obtain a better segmentation of images with many hairs.

The segmentation of images through the active contour model without edges (Chan-Vese) allowed to locate most images with regions close to the skin color and also very noisy images,such as the scalp. Using the same model, in the case of composed images of a large area of shadow or reflection inside the lesion area or over border, it was not possible to detect the edge of lesions in some of these images or they were not defined correctly.

Due to the lesions different size on the images, ie, according to their proximity in time of acquisition, the definition of the curve on the image, applying the Chan-Vese model, establishes the accuracy detectionof the lesions border. Better results are obtained with curves closer the lesion, thus the lesion is detected quickly and decreases the possibility of detecting false edges, such as those caused by reflections. Therefore, the definition of a curve of 140×140 for larger lesions and of 40×40 for small lesions is important since it will influence the results fordetected lesions. It can be considered another method to define a ideal curve for the segmentation of a specific lesion, such as using another technique for the establishment of an initial contour nearest the shape and size of each lesion.

The use of morphological filters gave a better definition of the lesion, because allowed the treatment of segmented images, with the presence of external noise, interior holes as well as irregular edges. For some images, the application of this filter, did not allowed the correct elimination of external noise neither large inner holes. This was not possible since it can not be defined an element structure too large because of the subsquent harm of the irregular border.

The method developed in this paper allowed the segmentation and definition of the edge of the skin lesions. Considering the visual analysis performed by a dermatologist, the method obtained 94.36% of accuracy.Through the Chan-Vese model, the methodology proposed showed promising results on the detection of skin lesions edge. Information obtained by this method can be available to the dermatologist in order to assist him in the diagnosis of skin lesions.

The assessments made by the dermatologist to evaluate the results of segmentation can influence the final results of the detection if they are evaluated by another experts. The proposed method may achieve more significant resultsof those presented in this study, with the improvement of these techniques used to solve the encountered problems, such as the heterogeneity of the base images.

Having in account the need to improve the encountered problems and the possibility of continued development of the method, the following tasks can be analyzed for the progression of the methods related to detection and classification of medical images: study or development of methods that treat reflections and shadows, without restricting the acquisition of images, allowing acquire images even by mobile devices, which have good image quality; consider a method to define a curve next to the lesion, so it can be detect quickly and accurately by use of the Chan-Vese model; analyze other segmentation technique such as the use of fuzzy logic to be compared with the Chan-Vese model; finally use other types of skin lesions.

ACKNOWLEDGMENT

The authors thank the "Conselho Nacional de Desenvolvimento Científico e Tecnologia - CNPq" by financial support. The authors are also grateful to Dr. Ricardo Rossetti Baccaro, Dermatologist Derm Clinic of São José do Rio Preto, which helped us in developing this paper related skin lesions.

REFERENCES

[1] Q. Abbas, I. Fondón, M. Rashid, "Unsupervised skin lesions border detection via two-dimensional image analysis," Computer Methods and Programs in Biomedicine, COMM-3090, pp. 1-15, 2010.

[2] M. Amico, M. Ferri, I. Stanganelli, "Qualitative asymmetry measure for melanoma detection," In: IEEE International Symposium on Biomedical Imaging: Nano to Macro, vol. 2, 2004, pp. 1155-1158, ISBN: 0-7803-8388-5.

[3] C. A. Z. Barcelos, M. Boaventura, E. C. Silva Junior, "A well-balanced flow equation for noise removal and edge detection," IEEE Transactions on Image Processing, vol. 12, n. 7, pp. 751-763, 2003.

[4] C. A. Z. Barcelos, V. B. Pires, "An automatic based nonlinear diffusion equations scheme for skin lesion segmentation," Applied Mathematics and Computation, v. 215, pp. 251-261, 2009.

[5] A. T. Beuren, R. J. G. Pinheiro, J. Facon, "Abordagem morfológica de segmentação do melanoma," In: Workshop de Visão Computacional, 7. Curitiba, 2011, pp. 249-254.

[6] Brasil. Ministério da Saúde. Instituto Nacional de Câncer, Estimativa 2010: Incidência de câncer no Brasil, Rio de Janeiro: INCA, 2009. 98 p.

[7] T. F. Chan, L. A. Vese, "Active contours without edges, IEEE Transactions on Image Processing," vol. 10, n. 2, p. 266-277, 2001.

[8] P. Cudek, J. W. Grzymala-Busse, Z. S. Hippe, "Melanocytic skin lesion image classification, Part I: Recognition of skin lesion," In: Conference on Human System Interactions (HSI). 3rd, Rzeszow, Poland, 2010, pp. 251-257.

[9] Dermatlas. B. A. Cohen, C. U. Lehmann, Johns Hopkins University - DermAtlas, Disponível em Dermatology Image Atlas: <http://dermatlas.med.jhmi.edu/> Acesso em: 2012.

[10] Dermatology Database. Y. Suzumnura. YSP Dermatology Image Database - Japan, Disponível em YSP Dermatology Image Database: <http://homepage1.nifty.com/ysh/indexe.html>. Acesso: em 2012.

[11] Dermis. Diepgen TL, Yihune G et al. Dermatology Information System - DermIS, Disponível em Atlas Dermatológico Online: <http://www.dermis.net/dermisroot/en/home/index.htm>. Acesso em: 2012.

[12] Dermnet, Skin Disease Atlas. J. L. Campbell Jr., Nevi, melanoma, Disponível em Malignant Melanoma e Melanocytic Nevi: <http://www.dermnet.com/videos/nevi-melanoma/>. Acesso: em 2012.

[13] Dermnet, Skin Disease Atlas. S. Chapman, Benign tumors. Disponível em Seborrheic Keratosis: < http://www.dermnet.com/videos/benign-tumors/>. Acesso: em 2012.

[14] J. Facon, Morfologia matemática: teoria e exemplos, Curitiba: Editora Universitária Champagnat da Pontífica Universidade Católica do Paraná, 1996. 320 p.

[15] J. L. Meded, Melton, & MD, Editores. Skin Cancer and Benign Tumor Image - Loyola University - Chicago, Disponível em Loyola University Dermatology Medical Education: <http://www.meddean.luc.edu/lumen/MedEd/medicine/>. Acesso em 2012.

[16] D. Mumford, J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," Communications on Pure and Applied Mathematics, v. XLII, pp. 577-685, 1989.

[17] K. –A. Norton, H. Iyatomi, M. E. Celebi, G. Schaefer, M. Tanaka, K. Ogawa, "Development of a novel border detection method for melanocytic and non-melanocytic dermoscopy images," In: Annual International Conference of the IEEE EMBS, 32nd. Buenos Aires, Argentina, 2010, pp. 5403-5406.

[18] S. Osher, J. A. Sethian, "Fronts propagating with curvature dependent speed: algorithms based on Hamilton-Jacobi formulations," Journal of Computational Physics, vol. 79, pp. 12-49, 1988.

[19] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-9, pp. 62-66, 1979.

[20] P. Perona, J. Malik, "Scale-space and edge detection using anisotropic diffusion", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, n. 7, pp. 629-639, 1990.

[21] R. B. Oliveira, "Método para detecção e classificação de lesões de pele em imagens digitais a partir do modelo Chan-Vese e máquina de vetor de suporte", 2012. 134 f. Dissertação de Mestrado em Ciência da Computação, Instituto de Biociência, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto, 2012.

[22] Saúde Total. Câncer da Pele: fotoproteção, Vida saudável com o sol. Disponível em <http://www.saudetotal.com.br/prevencao/topicos/default.asp>. Acesso em: 2012.

[23] Skin Cancer Guide, Melanoma, Disponível em <http://www.skincancerguide.ca/melanoma/images/melanoma_images.html>. Acesso em: 2012.

[24] N. Zhang, J. Zhang, R. Shi, "An Improved Chan-Vese model for medical image segmentation", In: International Conference on Computer Science and Software Engineering, Wuhan, Hubei, 2008, pp. 864-867.

[25] H. Zhou, G. Schaefer, M. E. Celebi, H. Iyatomi, K.-A. Norton, T. Liu, F. Lin, "Skin lesion segmentation using an improved snake model," In: Annual International Conference of the IEEE EMBS, Buenos Aires, Argentina, 2010, pp. 1974-1977.

SESSION 2

MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

A Decision Support System for Product Category Space Allocation in Retail Stores

Fábio Pinto
Faculdade de Engenharia
da Universidade do Porto
Rua Dr. Roberto Frias, s/n
Porto, Portugal 4200-465
Email: fabiohscpinto@gmail.com

***Abstract*—Maximizing the profitability of the available resources is a major task for retail companies. Space, as one of the most expensive resources, is one of the variables of more difficult allocation. This paper seeks to develop a Decision Support System to assist retailers in assigning space to product categories within a store. Through a combination of sales forecasting models and a Genetic Algorithm for optimization purposes, we access the quality of the developed system with a real case study. Results show that our system can successfully suggest space recommendations very similar to those of the business specialists.**

I. Introduction

The growing competitiveness in the retail industry is an evidence even for the most inattentive consumer. Retailers compete daily for the loyalty of their customers through diverse marketing actions while providing their stores with better products, better prices and better customer service. In this context of extreme competition, the ability to maximize the profitability of the available resources is crucial.

Knowledge Discovery in Databases (KDD) is a set of procedures with the aim of finding valid, novel, useful and understandable patterns in data. Data Mining (DM), as one of those procedures, focus on applying data analysis and discovery algorithms to extract knowledge from data [1]. Nowadays, several fields benefit from these techniques, including telecommunications, marketing, investments or manufacturing. Decision takers seek in a growing trend to run their businesses with a data driven approach. Retail, as one of the activities that more data generates, is also one of the industries that can benefit more with KDD.

Space, as one of the most expensive resources for retailers, is also one of the most difficult to manage. The number of product categories is growing and their allocation within a store layout[1] is becoming more difficult. KDD can support retailers in understanding the relation between space and sales of a product category and optimize store layout.

We developed a Decision Support System (DSS) that seeks to assist decision takers with the problem of assigning the optimal amount of space to a given product category within a retail store layout. Space elasticity is deeply related to this problem: the impact in a given product category sales by increasing a percentage (typically 1%) of that product category's space. At the moment of assigning space to the growing amount of product categories, it is essential for the retailer to access the effects of allocate more space to product category *x* instead of product category *y*.

Our approach is exposed in Figure 1. Firstly, we collected data that allowed us to access the given problem. It was

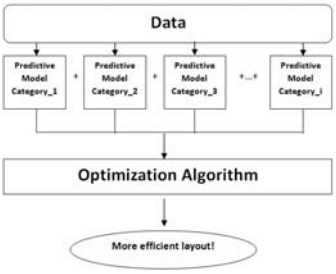


Fig. 1: Decision Support System architecture.

clear that the chosen variables would have a tremendous impact on the success of the project. Secondly, the collected data was used to develop forecasting models at the product category level. Each one of those predictors enabled us to model the relation between space and sales for product categories. Finally, the forecasting models were combined with a optimization algorithm (a Genetic Algorithm) that searched the set of areas (one for each product category, each forecasting model) that maximized sales. This set of areas is the final result of the DSS and aims to assist retailers in assigning space to the different product categories. The project was conducted in collaboration with a Portuguese retail company.

This paper is organised as follows. In Section II we describe the collected data and specify the data selection process for the first modelling phase. Section III exposes the methodology carried while building the sales forecasting models. Section IV and V describes two experiments in modelling space elasticity: first, only four carefully selected product categories were modelled; secondly, we chose to model all product categories. In Section VI we specify how we combined the forecasting models with a optimization algorithm for our DSS. Our case study is presented in Section VII in which we evaluate the quality of the developed DSS. Section VIII presents related work. Finally, in Section IX we present some conclusions and define future work. The project followed the CRISP-DM [2] methodology as Data Mining process model.

[1]Sketch showing the physical distribution, sizes and weights of elements such as text, graphics or pictures in a store.

II. DATA

This Section specifies the data collected for the project and the process of data selection for the first phase of space elasticity modelling.

Following the CRISP-DM methodology, the collected data passed through a cleaning and merging process that will not be specified here. Finished those processes, a data exploratory analysis was followed which, due to space restrictions, will not, again, be exposed in this paper. However, the exploratory analysis of the data gave us confidence about its quality and meaning for the project.

A. Data Collection

Table I specifies the variables collected for the project. The dataset comprises two years[2] of data with monthly observations for 110 product categories. Overall, the dataset brings together 332885 observations.

TABLE I: Collected data.

Variable	Description
Sales	Product category (PC) sales, within a store, at a given month.
Area_t	PC total area, within a store, at a given month.
Area_pe	PC permanent area, within a store, at a given month.
Area_pr	PC area, within a store, at a given month.
Month	Observation's month.
Insignia	Store's insignia.
Cluster	Store's sales potential cluster.
Cluster_Client	Store's client profile cluster.
PPI_County	Purchasing Power Index of the store's county.
N_W_Days	Number of non-working days of the month.
C_P_Index	Category penetration index by store's client profile cluster.

The majority of the variables were provided by the retail company in which this project was developed. We can not give insights about their construction but we can expose the purpose for their inclusion in our dataset:

Sales. The target variable. Reflects the liquid sales of a product category.

Area_t. An obvious inclusion. This variable represents the total area[3], in square meters, of a product category. [3] proved its significance on sales forecasting models for retail.

Area_pe. Represents the permanent area, in square meters, of a product category. Permanent area does not change due to seasonality factors. This type of area only changes during store layout restructuring.

Area_pr. Represents the promotional area, in square meters, of a product category. Promotional area changes due to seasonality factors.

Month. Included for seasonality purposes. It is common sense that retail sales are highly seasonal and we expected to model that volatility with this nominal variable.

Insignia. The retail company has three different types (insignias) of stores. This nominal variable attempted to capture different sales behaviour within each of these insignias.

Cluster. The retail company divides its stores in four distinct clusters according to their sales potential. Given the nature of the (nominal) variable, its inclusion in our dataset seemed relevant.

Cluster_Customers. The retail company divides its stores in four distinct clusters according to the profile of their customers. Again, the inclusion of this nominal variable seemed relevant given that it is expected that different customers will result in stores with different sales behaviour.

PPI_County. Purchasing Power Index of the county in which the store is located. It is expected that the bigger this variable, bigger the sales.

N_W_Days. Customers do the most of their shopping at non-working days so it is expected that the bigger this variable, bigger the sales.

C_P_Index. Discrete variable calculated for each product category within each customer cluster, so, there are 4 indexes by product category, one for each cluster. This variable can capture the impact that different customers have in product category sales.

B. Data Selection

Naturally, changes of space in a retail store does not occur monthly. Although some product categories undergo more space changes than other, the number of categories with static sales area is always high. Given that the main goal of the project is to implement an optimization algorithm on the variable *Area*, is very important to realize that models can only answer to circumstances for which data is supplied. This fact alerted us for the representativeness of the sample to the question that we want to answer.

Take the example of Figure 2: "Sample A", being *Area* the variable *x1* and *x2* any other independent variable, is better than "Sample B" (for these variables); in the same way as "Sample B" is better than the "Sample C". We believed that many product categories would resemble "Sample C". With this in mind, we searched the dataset for the product categories with greater space changes/variation. Those carefully selected product categories shall be the first categories to model in Section IV.

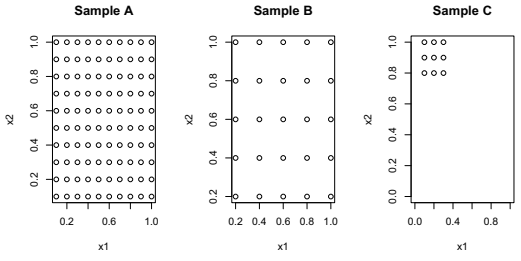


Fig. 2: Sample representativeness.

As measures of space volatility, two statistics by product category were introduced:

Number of mean changes (*NMC*) of the sales space,

[2]Between 2009 and 2010.

[3]Sum of permanent and promotional area.

defined by

$$NMC = \frac{\sum_{s=1}^{n} f_{m,s}}{n} \quad (1)$$

with

$$f_{m,s} = \begin{cases} 1 & \text{if } a_{m,s} \neq a_{m-1,s} \\ 0 & \text{otherwise} \end{cases}$$

where *s* is the store, *m* is the month and *a* is the variable *Area*.

Absolute mean monthly variation (*AMMV*),as percentage, of the sales space, defined by

$$AMMV = \frac{\sum_{s=1}^{n} \frac{\sum_{m=2}^{k} u_{m,s}}{k}}{n} \times 100 \quad (2)$$

with

$$u_{m,s} = \left| \frac{a_{m,s} - a_{m-1,s}}{a_{m-1,s}} \right| \quad (3)$$

where *s* is the store, *m* is the month and *a* is the variable *Area*.

The combination of NMC and AMMV gave birth to another measure of space volatility, defined as

$$score = NMC \times AMMV \quad (4)$$

this statistic shall be used to select the product categories with greater space volatility, which, we believed, were the product categories that could guarantee us the most reliable predictive models.

Figure 3 shows the values of NMC and AMMV registered for each product category that integrates our dataset. As we

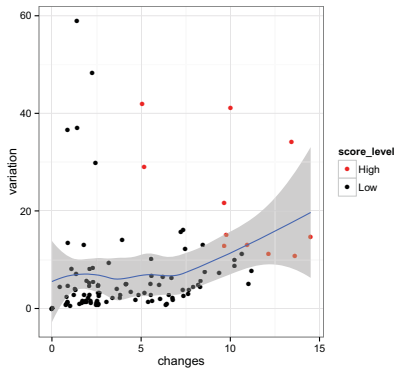


Fig. 3: Changes and variation of space.

suspected, a great portion of the product categories lie in lower left side of the graph, meaning that those categories have a reduced space volatility. The point colour of each product category is defined according to their *score* value[4]. From the *high score* group of product categories, we shall select three or four for a first modelling phase.

However, the product categories selection shall not be restrictedly based upon their space volatility. It is wishful that the selected categories present a homogeneous sales behaviour within the store population. The purpose of this requirement is to reduce the impact of the factor "store" in the relationship between product category space and sales. For that, the standard deviation of sales as a percentage of total sales by product category was calculated. The lower this value, the more homogeneous are the sales of the product category in the set of stores.

At this point, we already had the indicators that could lead us to select three or four product categories for a first phase of modelling. So, among the categories with *high score*, we sought for some diversity within this group: categories with a lot of volatility in the permanent area and little in the promotional area; categories with little volatility in the permanent area and a lot in the promotional area; and categories with a similar level of volatility in the permanent and promotional area.

The chart of Figure 4 presents the indicators for the eleven *high score* product categories: on the x-axis, the *score* for the permanent area of the categories; on the y-axis, the *score* for the promotional area; and the size of each dot relates to the sales homogeneity value of the respective product category, which are distinguished by different colours.

The selected categories are:

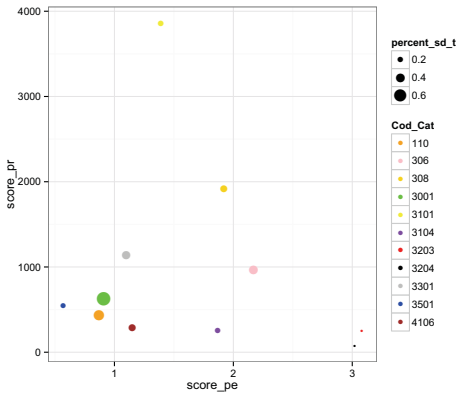


Fig. 4: Product category selection.

- **3204** - high volatility in terms of permanent area; sales homogeneity above-average.
- **3101** - high volatility in terms of promotional area; sales homogeneity slightly below average.
- **3001** - considerable volatility in both area components; is the product category with the best record of sales homogeneity among the categories whom *score* level was classified as *high*.
- **308** - considerable volatility in both area components; is the product category with the worst record of sales homogeneity among the categories whom *score* level was classified as *high*. It was expected that the results obtained for this product category should be worse than the results obtained for the others.

III. Methodology

This Section specifies the applied methodology for developing the predictive models. The Data Mining task carried in this project was regression and the dataset consisted in its

[4]If a category has a *score* value equal or greater than the 90th percentile of all values of *score*, that category is considered to have a *high score* level.

totality by temporally referenced observations. Both facts had implications in the chosen methodology.

A. Error Measures

The error measures for the evaluation of the predictive models are:

Mean Relative Error, defined by

$$MRE = \frac{\sum_{i=1}^{n} |\frac{x_{1,i}-x_{2,i}}{x_{1,i}}|}{n} \times 100$$

Root Mean Squared Error, defined by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_{2,i}-x_{1,i})^2}{n}}$$

$\mathbf{R^2}$, defined by

$$R^2 = 1 - \frac{\sum_i (x_{1,i}-x_{2,i})^2}{\sum_i (x_{1,i}-\bar{x})^2}$$

where, for all measures,

- $x_{1,i}$ is the real value;
- $x_{2,i}$ is the predicted value.
- $\bar{x}$ is the mean of the real values.

The performance of the regression algorithms shall be compared with two linear models: a simple linear regression; and a baseline model, whose predictions consist of the average of the target variable for the store which is being predicted. These comparisons will allow to access the difficulty of the phenomenon that we are modelling.

The retail company defined 10% as MRE target value. We assumed this value as an indicator of the predictive capacity of our models.

B. Error Estimation Methodology

Once the dataset used in this work consisted of multiple time series, it is crucial to apply a methodology that takes into account the temporal reference of the observations.

The *sliding window* [4] method adds two important features to the classic *holdout* method: 1) takes into account the temporal order of observations that constitute the training and test set, meaning that all observations which form the training set are prior to observations that form the test set, simulating the forecasting process that occurs when using the models; 2) considers possible regime changes in the phenomenon that one is trying to predict, since it is possible to change the training set "window", eliminating older observations and adding the most recent ones, retaining the training set a constant dimension.

C. Regression Models

Several regression models were tested, namely:

- **Cubist**, an improvement of Quinlan's M5 regression tree [5].
- **Artificial Neural Networks** (ANN), following the empirical evidence of the capacity of ANN to successfully predict retail sales [6].
- **Multivariate Adaptive Regression Splines** (MARS), as exposed in [7].
- **Support Vector Machines** (SVM), as exposed in [8].
- **Generalized Boosted Models** (GBM), an *R* package for boosting [9] models.
- **Random Forests** (RF), Breiman's bagging of trees [10].

All regression models were fitted through *R*, software for statistical computing, and several *R* packages.

IV. Modelling Space Elasticity: Selected Product Categories

This Section exposes the experimental setup, results and discussion in modelling the four product categories selected in II-B.

A. Experimental Setup

The training set for each product category (and each model) consisted in one and a half year of observations; the remaining examples (half a year) are reserved for the test set. All generated models follow the variable specification shown in Table II. During the modelling experimental phase, several combinations of variables were tested. However, the inclusion of the variable $Sales_{t-1}$ had a great impact on predictive performance.

TABLE II: Models variables.

Variable	Description	Variable type
Sales	Product category sales.	Dependent variable.
Area_t	Product category total area.	Independent variable.
Month	Observation's month.	Independent variable.
Insignia	Store's insignia.	Independent variable.
Cluster	Store's sales potential cluster.	Independent variable.
Cluster_Client	Store's client profile cluster.	Independent variable.
PPI_County	Purchasing Power Index of the store's county.	Independent variable.
N_W_Days	Non-working days of the month.	Independent variable.
C_P_Index	Category penetration index by store's client profile cluster.	Independent variable.
$Sales_{t-1}$	Product category sales at t-1.	Independent variable.

The tuning of the models was done on the test set, assisted by the *R* package *caret*.

B. Results

Table III summarizes the results obtained while modelling the four selected product categories.

TABLE III: Selected product categories results.

Product Categories		3204	3101	3001	308
Algorithm	Error Measure				
Cubist	MRE	90.91%	28.09%	21.95%	16.58%
	RMSE	246.97	3065.61	15545.41	5827.96
	R^2	0.44	0.92	0.97	0.97
ANN	MRE	106.18%	25.41%	27.51%	17.27%
	RMSE	193.54	2187.17	22565.20	5421.74
	R^2	0.42	0.96	0.96	0.97
MARS	MRE	113.47%	27.98%	44.50%	24.22%
	RMSE	276.27	2458.29	16537.03	6744.37
	R^2	0.40	0.95	0.96	0.95
SVM	MRE	61.89%	16.70%	22.42%	12.96%
	RMSE	184.35	1977.63	15905.20	5060.28
	R^2	0.56	0.97	0.97	0.98
GBM	MRE	82.01%	83.73%	29.88%	25.71%
	RMSE	207.20	5512.66	24973.59	12347.00
	R^2	0.47	0.79	0.91	0.85
Random Forest	MRE	76.25%	24.36%	21.17%	12.44%
	RMSE	165.18	2650.14	15851.46	6895.59
	R^2	0.57	0.94	0.97	0.97
Linear Regression	MRE	86.90%	174.61%	121.36%	70.12%
	RMSE	219.89	4564.30	26619.49	16484.80
	R^2	0.49	0.82	0.89	0.70
Baseline	MRE	59.34%	46.90%	143.73%	29.79%
	RMSE	142.19	5462.56	65278.39	22247.79
	R^2	0.69	0.72	0.34	0.48

C. Discussion

The *MRE* estimated for the models is far from the target value of 10%, except for product category **308**. This category already presented models with reasonable predictive accuracy. However, for the other product categories, the error is higher. In fact, for one product category (**3204**), the *baseline* model had a lower error than the fitted regression models. This is a clear indicator that the models generated for that particular product category are bad.

Comparing these results with the assumptions leading to the selection of the four product categories, it is clear that there is a large discrepancy. There seems to be no correlation between the variables calculated to characterize the product categories and the predictive performance of the algorithms. This result is partially explained in [11]. Nevertheless, this first experiment gave important information for the modelling of all product categories, exposed in Section V.

V. Modelling Space Elasticity: All Product Categories

Given that the results obtained in Section IV were not satisfactory, we moved to model all the product categories and evaluate its results. This section exposes the setup, results and discussion for that experiment.

A. Experimental Setup

Analysing the results of Table III, two algorithms appear to have the most robust performance: Random Forest and Support Vector Machines. However, there is one important detail about the SVM model of the category **3204**: its *sigmoid kernel* function, opposing the *radial kernel* function of the other three SVM models generated so far. This fact gains extraordinary importance given that this particular version of SVM shows much better performance than the other regression models.

Some of the 110 product categories showed a low number of observations. We set the minimum of observations for modelling purposes at 1000. This conditioning resulted in 89 product categories that we were able to model.

So, three models ere fitted to the filtered dataset:

- **Random Forest**
- **Support Vector Machine**, with *radial kernel* function.
- **Support Vector Machine**, with *sigmoid kernel* function.

Again, we followed the methodology exposed in Section III and the models were fitted including the variables explicit in Table II. The parameter tuning of the models was done in the test set.

B. Results

Figure 5 shows the results obtained for the three fitted models in terms of MRE, for 89 product categories. The black horizontal line represents the target error defined in Section III. Several models are below or very close that line. This is a good indicator that the results obtained from modelling all product categories are far better than those obtained modelling only the four selected categories.

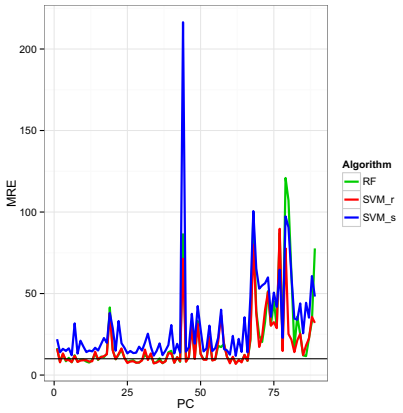


Fig. 5: Results for DM models (MRE)

In Table IV we compare the best DM model for each product category with the respective baseline model. On average, DM models show a better predictive performance than the baseline models.

TABLE IV: Summary: best DM model *vs* Baseline (MRE)

Model	Min	1stQ	Median	Mean	3rdQ	Max
Best DM	7,20	8,66	11,62	17,60	20,29	86,38
Baseline	7,68	10,58	13,78	27,73	27,83	198,90

C. Discussion

Overall, Random Forests obtained the best performance[5] in 41 product categories; SVM with *radial kernel* function

[5] Algorithm's performance evaluation was done using the three error measures defined in Section III.

was better in 45 product categories; and finally, the SVM with *sigmoid kernel* function presented superior performance in 3 product categories. In several product categories, Random Forests and SVM with radial kernel function showed a very similar performance. Here, we choose SVM instead of Random Forests, given that its computationally less demanding.

Although some models presented high error, half of them had a performance very close or better than the target value set by the retail company, as the median of the error in Table IV proves.

VI. Decision Support System

Once the predictive models were generated, we were in conditions to combine these with an optimization algorithm in order to develop the DSS for product category space allocation. The applied optimization algorithm was an adapted version of a Genetic Algorithm (GA). This Section presents a brief introduction to GA and Evolutionary Computing, followed by the specification of the GA applied in our DSS. Finally, the Section ends with an application and evaluation of the developed DSS.

A. Genetic Algorithms

Evolutionary Computing [12] is divided into four sub-branches: Evolutionary Programming, Evolutionary Strategies, Genetic Programming and Genetic Algorithms. Figure 6 illustrates a Evolutionary Computing algorithm scheme.

It was Holland that proposed GA [13] in the 60s/70s. As the name suggests, Holland was inspired by the Darwin's Theory of Evolution[6] to develop these algorithms, whose primary objective was to study adaptive behaviours.

A GA comprises five key stages: defining the solutions representation; parents selection; recombination of solutions (crossover), mutation, and survival selection.

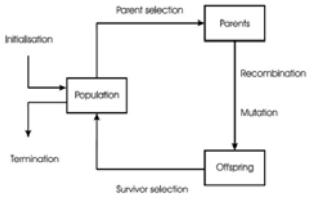


Fig. 6: Evolutionary Computing. Adapted from [12].

In the classic version of the GA, the solutions representation is binary, which translates into a certain code that represents a solution to the problem which is being optimized. Each element of a solution is called a "gene". The initial population of solutions is typically generated randomly, followed by a selection based on probabilities (all solutions have a chance to be selected). The selection probability estimation of each solution is based on the fitness function defined (the higher the fitness of the solution is, the higher the probability of being selected for crossover is) which evaluates each solution and returns the respective output. This group of selected solutions is called *parents*.

The crossover operator allows the combination of the *parents* solutions and from that new solutions are generated, called *children*. Crossover occurs through the random selection of two elements of *parents* and point of partition in both. The new solution arises from combining the opposing parties of the *parents*. Then follows the mutation process, which through a mutation rate, randomly selects one or more solutions of the *children* group to be mutated. This mutation occurs through the selection of a binary element of the selected *children* and its "flip", i.e., if the element is 0 becomes 1 and vice versa.

Generated the *children* group, follows the replacement of the initial population by the children group. The fitness function evaluates again all solutions of the population and the cycle restarts. The process ends when a certain condition is met, which may be a number of iterations or a particular value for the best fitness solution.

B. Specification of the DSS

It is essential that the DSS development considered the specific issues concerning the phenomenon that we were optimizing. The solutions that could arise from the process must not violate three fundamental restrictions:

- a minimum value for each "gene" (the minimum space of a certain product category)
- a maximum value for each "gene" (the maximum space of a certain product category)
- a maximum value for the sum of all "genes" (the maximum space of the sum of all product categories and layout)

Solutions representation. A solution is made by the area of each product category on a given store at a given time. Each solution will have as many elements as the number of product categories in the layout that is being optimized. The initial population consists of 30 solutions, 10 of which are the homologous area of the time that is being optimized and the remaining 20 are random solutions generated within the restriction set[7]. The fitness function that evaluates the solutions consists of the predictive models generated in Section V, that through the input of the solutions output sales forecast. The sum of all sales forecasts in all categories present in the solution constitutes the fitness value of that solution.

Parents selection. The selection of the solutions that will generate new solutions is based in a probability associated with each, taking into account the fitness function output for each solution:

$$Prob(x) = \frac{Fitness(x)}{\sum Fitness} \tag{5}$$

Then, from 30 solutions that constitute the population at this point, 10 are selected for crossover.

Crossover. In this stage, the 10 selected solutions, generate 10 new solutions. In order to not violate the restrictions that were imposed on all solutions, we need a specific crossover

[6]Charles Darwin, author of "On the Origin of Species", published in 1859.

[7]The maximum and minimum space restriction for each product category is defined by the respective maximum and minimum values of space that occurred in the most recent year available in the dataset for the store that is being optimized.

operator. The answer is given by an operator that, given two solutions selected for crossover, calculates the mean of the two, originating a new solution. This procedure allows to inviolate the conditions mentioned above.

Mutation. Given a mutation rate, this operator proceeds to randomly select a subset of solutions from the *children* group. Two elements are randomly chosen from each of the selected solutions, to which is added (to the first) and subtracted (to the second) the value *m*. This value is calculated by averaging the two elements of the solutions selected for mutation divided by 100, thus ensuring that the overall layout space is not exceeded. However, this operator may disregard restrictions on the minimum and maximum area of each product category. In order to circumvent the problem, the solution is corrected if one of the mutated elements is outside the defined range, by assigning the closest border value[8].

Survival Selection. Completed the crossover and mutation processes, the 10 new generated solutions are added to the solution population. With a total population of 40 solutions, a new evaluation by the fitness function occurs. The solutions are then sorted regarding the respective output of the fitness function, and the top 30 remain for new iteration. Here, the cycle restarts.

The DSS performs optimization at the monthly level: given a store, a particular number of product categories and one month, the algorithm seeks the set of areas that maximizes monthly sales according to the predictive models generated in Section V.

VII. Case Study

Developed the predictive models and the DSS that combined them with an optimization algorithm, we were in conditions to evaluate the quality of our system. That opportunity was proposed by the retail company in which this project took place.

In May 2011, the store *X* had a makeover of its logistics which included a new layout. The vast majority of the product categories that were available in this store suffered space changes. For this makeover, the analysts of the retail company based their space recommendations through analysing data from 2009 to 2010, the data that allowed us to construct our dataset. Thus, it was proposed to test the developed DSS in this store and compare the results obtained with the recommendations of the business specialists.

A. Experimental Setup

As assumptions for this experiment, we assumed that, except the variables *Area_t* and *N_W_Days*, the other independent variables remained constant, given that seems acceptable that from 2010 to 2011 they have not changed substantially. The predictive models generated for this experiment integrated all observations available in the dataset. As had they been previously tested, we do not admit the need to set aside a test set. To generate the initial population of solutions, the real areas of the store in 2010 were used.

Of the 89 categories for which it is possible to build predictive models, 85 are part of the store layout, so, the DSS had to optimize 85 values. To evaluate the quality of the optimization and compare its results with recommendations from the business specialists, we used R^2 as defined in Section III.

B. Results

For 50000 iterations, the optimization algorithm presents the performance[9] showed in Figure 7. It is observable a slower growth of the maximum fitness value starting at 40000 iterations. The chart exposes the capacity of the designed optimization algorithm in seeking solutions that maximize the output of the fitness function under the defined restrictions set.

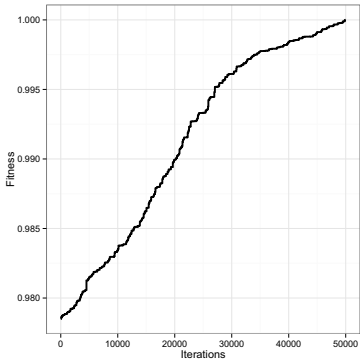


Fig. 7: DSS results for 50000 iterations, max fitness

Comparing the results obtained by the DSS with the space recommendations made by the business specialists, we estimated $R^2 = 0.70$. Through several experiments, we accessed the robustness of this result. Given the "leave-one-out" nature of the experiment and assuming that the space recommendations of the business specialists are ideal, this results made us optimistic about the quality of the developed DSS.

C. Discussion

Analysing the set of areas in detail, arise some important questions. A first comparison between the output of the DSS and the real areas of the store *X* at May 2011, three product categories are clearly overvalued by the system. These categories have a common factor: they receive at specific times of the year very strong promotional campaigns. It is clear that the models of these product categories (with MRE 22.44%, 11.82% and 21.77%) are failing to associate this boom in sales to seasonal factors. Those strong promotional campaigns also imply space (promotional area) increase, which mistakes the predictive models for the reasons behind the sales boom. Given that the *Month* variable varies systematically and the

[8]Example: if the minimum value of the product category in question is 3, and the mutated element is 1, occurs the substitution of one by another.

[9]Due to confidentially purposes, the fitness values scaled between 0 and 1, with 0 being the minimum fitness value and 1 being the maximum fitness achieved by the system.

Area_t presents a great increase at those times (just like the *Sales* value), the forecasting models assigned the cause of the sales boom to the *Area_t* variable. This has implications in the DSS output: for these product categories, its recommendation is to have promotional campaigns at any time of the year.

However, there are some interesting points in these results. For the adjacent product categories 1502 and 1503, the DSS suggests areas quite distinct from the recommendations of the business specialists. Given the excellent predictive accuracy of the models in question (7.20% and 10.79%, respectively), we can consider this result as useful information for the retail company.

VIII. Related Work

Space allocation within a retail store is a common research topic. Several studies were conducted for studying the process of product allocation using Econometrics [14], Operations Research [15] and even Machine Learning [16]. However, these works do not concern product category allocation. They focus on which products should have exposure in the retail store. We focus on the problem of the amount of space that the category of that particular product should have within the retail store.

Desmet and Renaudin [17] published the first work concerning the problem of product category space allocation. They used Econometrics to model sales behaviour of product categories and estimated the respective space elasticity for each. Given the linear nature of the models fitted, they could not apply any kind of optimization technique and their results were partially questionable, with some estimated space elasticities being negative values.

Castro [3] followed an approach very similar to [17]. Again, the linear nature of models did not allow an optimization approach. Space elasticities were estimated for each product category and its results were an important contribution for this work.

Given the bibliographic research that has been done, we believe that this is the first work of this nature to be published. We hope that our contribution can bring more attention for the problematic and promote more work under the topic.

IX. Conclusion and Future Work

This paper exposes the combination of sales forecasting models and a Genetic Algorithm to develop a DSS for product category space allocation in retail stores.

Once the collected data could not have the necessary granularity for the modelling purposes, our first approach to the dataset was to carefully select the product categories that suited better our assumptions. Although we successfully found those product categories, the results of modelling those categories were not the expected. However, in the second modelling experiment, the results were far better and the forecasting models presented a good predictive performance to the quality standards of the retail company in which this project took place.

The sales forecasting models were successfully combined with a Genetic Algorithm, specifically adapted to the optimization constraints. This combination results in a system that can recommend space allocations in retail stores very similar to the recommendations made by business specialists. We described a case study in which our system showed a good performance.

As future work, while building the dataset that originated this paper, the need for three specific variables raised:

- *Promotional activities.* Variable that captures the level of a promotional activity within a product category, in terms of advertising, prices, etc. This variable could solve the overvalue of some product categories by the DSS.

- *Space quality.* Variable that captures the quality of the area assigned to a product category, i.e., if a product category is situated within the entrance of the store, it has more visibility.

Thus, for some product categories, the number of examples was low. This problem can be resolved by collecting more data. Ideally, this new data would reflect a systematic variation of product category space within a store(s), allowing for a sample more representative of the phenomenon.

References

[1] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, no. 3, pp. 37–54, 1996.

[2] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, *Cross-Industry Standard Process for Data Mining 1.0: Step-by-step Data Mining Guide*, SPSS Inc., 2000.

[3] A. Castro, "As vendas e o espaço no retalho: modelos econométricos aplicados a um grupo de distribuição alimentar português," Master's thesis, Faculdade de Economia do Porto, 2007.

[4] M. Datar, A. Gionis, P. Indyk, and R. Motwani, "Maintaining stream statistics over sliding windows," *SIAM Journal on Computing*, vol. 31, no. 6, pp. 1794–1813, 2002.

[5] J. Quinlan, "Learning with continuous classes," in *AI92*, Adams and Sterling, Eds., Singapore, 1992, pp. 343–348.

[6] I. Alon, M. Qi, and R. Sadowski, "Forecasting aggregate retail sales: a comparison of artificial neural networks and traditional methods," *Journal of Retailing and Consumer Services*, no. 8, pp. 147–156, 2008.

[7] J. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–141, 1991.

[8] V. Vapnik and C. Cortes, "Support-vector networks," *Machine Learning*, no. 20, pp. 273–297, 1995.

[9] J. Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[11] F. Pinto and C. Soares, "Predicting the accuracy of regression models in the retail industry," in *5th Planning To Learn Workshop WS28 at ECAI 2012*, 2012, p. 28.

[12] A. Eiben and J. Simth, *Introduction to Evolutionary Computing*, 1st ed., ser. Natural Computing Series. Springer, 2003.

[13] J. Holland, *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975, vol. Ann Arbor, no. 53.

[14] V. Gaur, M. Fisher, and A. Raman, "An econometric analysis of inventory turnover performance in retail stores," *Management Sci*, vol. 51, pp. 181–193, 2005.

[15] X. Dréze, S. J. Hoch, and M. E. Purk, "Shelf management and space elasticity," *Journal of Retailing*, vol. 70, no. 4, pp. 301–326, 1994.

[16] H. Hwang, B. Choi, and G. Lee, "A genetic algorithm approach to an integrated problem of shelf space design and item allocation," *Computers and Industrial Engineering*, no. 56, pp. 809–820, 2009.

[17] P. Desmet and V. Renaudin, "Estimation of product category sales responsiveness to allocated shelf space," *International Journal of Research in Marketing*, no. 15, pp. 443–457, 1998.

Measuring the Improvement of Web Page Classification in Using Mark-up Features

Pedro Strecht
Faculty of Engineering of the University of Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
Email: pstrecht@fe.up.pt

***Abstract*—This paper explores the benefits of using mark-up specific features to improve web page classification. Several experiments were conducted, using the Support Vector Machines algorithm, such that this kind of features are included incrementally. From these it is possible to show empirical results which allow measurements and comparisons to be made. Moreover, measuring the quality of web page classification leads to the identification of possible key features that have more impact in improving it.**

I. Introduction

Being a free environment and instantly available, the Web has become a very large repository of data and often the key resource used for finding information. In natural language processing terminology, a corpus is a set of texts, and several corpus make a corpora [1]. Such concepts can also be used in web pages. In fact, it has been reasoned that the whole Web can be perceived as corpora [2], an idea reinforced by projects such as WebCorp [3]. As the Web corpora became accessible, web page classification research emerged. Text categorization is the task of automatically arranging a set of documents into categories (classes or topics) from a predetermined set [4]. Likewise, web page classification is the process of assigning one or more predefined categories to a web page [5].

Both problems of text and web page classification are still under researched and several approaches have been proposed. Some use text classification methods in web page classification [6], while others point out that the latter is not just another case of the former [7]. The rationale is that a web page, unlike plain text, offers mark-up which, besides enforcing structure to the text, is a potential source of meta-information. This means that from mark-up several web page specific features can be extracted, thus, creating methods particularly tailored for web pages. Web page classification is, therefore, a problem still open for further research, particularly, in the subject of feature selection.

This paper explores the potential benefits of including mark-up specific features by evaluating its impact on the performance of web page classification (instead of using only plain text features). The remainder of this paper is structured as follows. Section 2 presents the sub-problems, main applications and a description of the state-of-the-art approaches to automating web page classification. Section 3 presents related work. Section 4 presents the methodology. Section 5 presents results and analysis. Section 6 presents the conclusion.

II. Web page classification

A. Sub-problems in web page classification

When designing an approach to web page classification, usually one has to consider the various sub-problems and select a suitable approach for each. Table I summarizes the sub-problems [5].

TABLE I
Sub-problems in web page classification.

Sub-problem	Alternatives
Point of view	Subject classification Functional classification Sentiment classification
Number of available categories	Binary classification Multi-category classification
Number of assigned categories	Single-label classification Multi-label classification
Type of assignment	Hard classification Soft classification
Category organization	Flat classification Hierarchical classification

Regarding the point of view, in subject classification categories encompass the topic of a web page, however, in functional classification they are related to its purpose. In sentiment classification, categories are about the opinions presented in a web page. Concerning the number of available categories, classification may be in two categories only or in several categories. On the number of categories that can be assigned to a web page, in single-label classification only one category is allowed while in multi-label classification two or more categories are allowed simultaneously. Considering the type of assignment, in hard classification, a web page can either be or not be of a category, whereas in soft classification a probability is associated with each category. Finally, regarding the way categories are organized, in flat classification categories are equally parallel while in hierarchical classification each category can have a number of subcategories.

B. Applications of web classification

The main motivation to classify a web page into one or more categories is assisting information retrieval tasks. Some examples include helping categorization of web directories such as dmoz [8] (web directories are collections of links

which provide a guided way to explore the web), improving quality of searches either in the search engines by providing focused domain-specific crawls [9] instead of a full crawl or in presenting clustered results ordered by rank of relevance [5]. Other applications include question answering systems (by means of question classification), web content filtering, assisted web browsing and knowledge based construction. Web page classification can be also used to classify an entire web site [5].

C. Automating web page classification

A web page is made of mark-up and text. From these, features can be extracted which become the information needed to classify it (either from a subject, functional or sentiment point of view). A recurrent problem in web page classification is to find the best set of features to collect from a web page. Text classification usually relies on the absolute frequency of each distinct word stem [10] and n-grams [1]. However, by exploring the nature of a web page, more features emerge that come from mark-up, structure, URLs and connections to others pages [5] (either explicitly with hyper links or implicitly with methods for identifying implicit links). These are called on-page features. Moreover, the same features on neighbours pages should also be considered (particularly siblings [11]), especially in cases where the web page being classified does not contain text [5].

Human classification of a web page is still the most accurate method of classification available. However, it is not feasible in a large corpus, since it is a time consuming process [12]. Therefore, an automated approach is mandatory as machines are much faster at collecting features from a web page than humans. A large number of algorithms of machine learning have been developed on the problem of automated classification (in several contexts) that can be used in the particular context of web page classification. Fundamentally two different approaches have been proposed in the literature: unsupervised and supervised. Both of them require human assistance.

In unsupervised methods, the goal is to identify possible categories for a set of data. This is done by searching for patterns and structure among all features available. The most common method for this approach is clustering, in which similar web pages are grouped together to form clusters. Although these suggest potential categories, human intervention is still needed to name them. To achieve maximum effectiveness, a large and varied web corpus should be used in order not to bias the categories to a particular topic. A very common and simple algorithm for clustering is K-means [13]. Yet, it requires previous identification of how many clusters should be identified, which may not be an easy decision.

In supervised methods, the goal is to design a predictive system that can categorize a web page, in one or more categories, within a certain degree of accuracy. In these algorithms, the classes are predetermined [14] and come from a finite set, previously made by a human. A training set of data is hand-labelled with these classifications. The algorithm task is to induce a mathematical model from the available features in the training set of data in order to show a relationship to the classification given to each web page. The model is then tested in a different set of data. Its predictive performance is evaluated by how close to the classification previously hand-labelled the automatically given one is. There are several supervised learning algorithms which differ in the approach used in analysing data. Examples include Naive Bayes, k-nearest neighbor (k-NN) and Support Vector Machines (SVM), all of which have been used in text classification [15].

III. Related work

Numerous studies on feature selection in web page classification have been carried out using different approaches. Riboni [16] conducted experiments using Kernel Perceptron and Naive Bayes classifiers, observing that the combination of the usual representation of web pages using local words with a hyper textual one can improve classification performance. Shibu et al [17] used web page summaries concluding that both Naive Bayes and SVM improve their respective performances by solely analysing the text on the web page. Ozel [18] developed a genetic algorithm that determines the best features for a given set of web pages, arguing that when features selected by his genetic algorithm are used and a k-NN classifier is employed, the accuracy of classification improves up to 96%. Selvakuberan et al [19] propose a method called Combined Feature Selection and Classification for effective categorization of web pages. Experimental results show that their approach improves the classification accuracy with the optimum number of attributes using four learning classifiers. Mangai et al [20] proposed a framework for feature reduction and state that performs better than most of the other feature selection methods.

IV. Methodology

Concerning the alternatives in Table I, this study focuses on functional point of view, multi-category, single-label, hard, flat automatic classification of web pages.

To carry out the experiences it was specifically developed a web interface system (using Apache 2.4.2, PHP 5.4.6 and MySQL 5.5.27). Its architecture is presented in Figure 1.

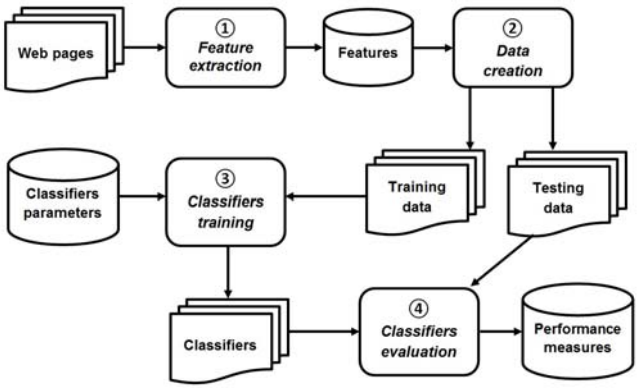


Fig. 1. System architecture.

The architecture for the system includes four distinct processes. In the first process, features are selected and measured from the web pages data set. All features are collected and stored in a relational database which facilitates process separation. Then for each type of features, a second process creates the corresponding training and testing data. After this, a third process creates classifiers for each training data using different combinations of parameters previously loaded in the database. Finally, a fourth process evaluates the performance of each classifier using the testing data and stores the results in the database. The evaluation measures are then compared enabling the identification of potential improvements in classification.

A. Data Set

The data set used is the "The 4 Universities Data Set" [21] It contains web pages gathered from the computer science departments of universities in the United States of America in January 1997, by the World Wide Knowledge Base project (CMU text learning group) [22]. It consists of 8282 pages that were manually classified into the following categories: course, department, faculty, project, other, staff and student. These categories are functional classification oriented, as they relate to the purpose of each web page. About 45% of the web pages belong to the "others" category. These web pages were dropped from the initial data set as this kind of label is not useful for the classification purposes of this study. This leaves a total of 4518 web pages for analysis. The category distribution is presented in Figure 2.

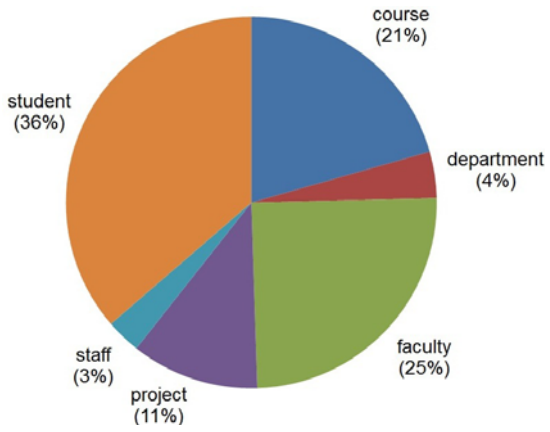


Fig. 2. Web pages category distribution.

It is worth noting that the web pages category distribution is quite unbalanced. The "student" category has 36% of the web pages, while the "department" and "staff" categories are the least represented, with only 4% and 3% of the web pages respectively.

B. Feature extraction

The feature extraction process includes the following steps for each web page of the data set:

1) Pre-processing of the content of the web page. This was done by converting all characters to lower case, removing meta-information, *Hyper Text Markup Language* (HTML) tags, extra-spaces and filtering all non-alphabetic characters.
2) Getting the bag-of-words representation. All words from the pre-processed content are extracted. Each word corresponds to a feature, however, it is useful to apply some basic feature reduction techniques such as:
   a) *Stemming*, in which all related words based on their root form are considered to be the same feature (stem). This was carried out using a PHP implementation of the Porter Stemming Algorithm [23] by Jon Abernathy [24].
   b) Filtering of *stop-words*, which are words so frequent in any text that their inclusion has no semantic meaning for classification. This was done by rejecting words included in an English stop-word list [25].
3) Calculating the term frequency. The most straightforward method to assign the weight of a term (is this case each stem) in a web page is to count the number of occurrences. This method is called *term frequency* (TF) [26] and is defined by the function:

$$TF(t_i, p_j) = \#(t_i, p_j) \quad (1)$$

   in which $\#(t_i, p_j)$ denotes the number of occurrences of term t_i in web page p_j.
4) Calculating the structured-oriented term frequency. The concept of term frequency does not account for mark-up in the web page. To meet the purpose of this study it is necessary to extract features which include structural information in the web page. One way to achieve this is to consider not only the number of occurrences of terms in web pages but also the HTML element in which they occur. Elements such as titles, headers and text enhancers (bold, italic and underline) provide context. This translates to rising the importance of terms occurring inside these elements. Therefore, there should be an increase in the weight given to these terms, as a function of the element they occur in. This approach was introduced by Riboni [16] from which the *structured-oriented term frequency* concept can be defined by the function:

$$SWTF_w(t_i, p_j) = \sum_{k=1}^{n} [w(e_k) \cdot TF(t_i, e_k, p_j)] \quad (2)$$

   where e_k is an HTML element, $w(e_k)$ is the weight assigned to element e_k and $TF(t_i, e_k, p_j)$ denotes the number of occurrences of term t_i within the context of element e_k in web page p_j and n is the number elements that are included in the analysis. It should be pointed out that the web pages in the data set do not contain meta information, so it is not useful to extract features from the `META` elements.
   In the study carried out by Riboni [16], several weights were tried for the `TITLE` element, ranging from 2 to 6.

In this study, this element has a weight of 5 and other elements receive different weights, which are presented in Table II. The choice of these elements relies on its wide use in the web pages of the data set.

TABLE II
WEIGHTS GIVEN TO WORDS BY ELEMENT OF OCCURRENCE.

Element of occurrence	Weight
TITLE	5
H1, H2	4
H3, H4	3
A,B,I,U,LI	2
All others	1

In the data set there are web pages that are not HTML well-formed. Examples of this are web pages where a <H1> tag is paired with a </H2> tag (possibly causing ambiguity to which is the element to be considered). The parsing of HTML was done using the PHP DOMDocument class which is robust enough to tolerate this kind of problems. It considers the opening tag as the one identifying the element. In the case of nested elements, the weight is given as the sum of all the elements in which the term appears (e.g. <H1><I>word</I></H1> would have a overall weight of 6). For these calculations, the same considerations for pre-processing are valid, although HTML tags were obviously not removed. Stemming and stop-words techniques were also applied.

5) Storing data about the web pages and the weight of their features in a relational database.

C. Train and test data creation

After all features are collected it is necessary to create training and testing data. Training data is used to build a classifier and testing data is used to evaluate its performance. To be possible to measure improvement in classification, mark-up features need to be added incrementally. For each level of increment a type of features is defined and a corresponding pair of training and testing data has to be created. Table III shows the training and testing data created for each type of features.

TABLE III
TRAINING/TESTING DATA CREATED.

Type of features	Elements analysed	$\#F$	FRR
0	All	42020	-
1	TITLE, H1,H2,H3,H4	9822	76,7%
2	TITLE, H1,H2,H3,H4,A	24108	42,6%
3	TITLE, H1,H2,H3,H4,A,B,I,U	25367	39.6%
4	TITLE, H1,H2,H3,H4,A,B,I,U,LI	31147	25,9%

The *Feature Reduction Rate* (FRR) is as a feature reduction indicator [27]. It compares the number of features in each type of features ($\#F_i$) to the ones in type 0 ($\#F_0$).

$$FRR_i = \frac{\#F_0 - \#F_i}{\#F_0} \quad (3)$$

As more elements are included in the analysis, the number of features increases causing the FRR to decrease.

It is also useful to measure improvement in web page classification of each type of features combined with type 0. This creates four additional sets of training/testing data (types of features 5 to 8 respectively). The process for the creation of training and testing data encompasses the following steps for each type of features:

1) Coding of categories. The six categories were coded with each receiving a code of 1 to 6. This was done only once.
2) Coding of features. Each feature in all the feature space was coded with an integer value. The codes vary for each type of features as the number of features is different.
3) Creation of data file. Each web page is represented as an array of features codes and corresponding weights. It is also included the category code it belongs to.
4) Scaling of the data file. Features weights are normalized to a 0 to 1 scale as there is evidence that it improves the performance of classifiers [28].
5) Shuffling of the data file. As the web pages in the data file are ordered by category code, it is necessary to randomize the order of the lines in the data file. This guarantees a fair distribution of web pages from all categories in both training and testing data.
6) Creation of training and testing data. From the shuffled data file, 70% of the lines are used for the training data file and the 30% remaining lines for the testing data file.

D. Classifiers training

To meet the purpose of this study it is necessary to perform several experiments in classifying web pages for each type of features. The classification algorithm chosen was the *Support Vector Machine* (SVM). The approach in the SVM algorithm is to define a vector space (each web page is a vector and each axis a feature) and to find a linearly separable decision space that best separates the vectors in two categories through a hyper plane. SVM produces very complex models which usually create new features in higher dimension spaces. The algorithm has the important property of its performance being independent of the number of features. It was demonstrated that SVM performs better than other algorithms in text categorization [15], thus making it a good choice for web page classification. Although initially designed for binary classification, SVM can also be used to multi-category problems such as the one concerning this study.

To create the SVM classifiers it was used the LIBSVM 3.14 system [29]. Several classifiers where trained for each type of features and kernel type. LIBSVM 3.14 uses the kernel types presented in Table IV.

For each kernel type, different parameters were used:

1) *Linear kernel*. The only available parameter is the soft margin constant (C), which took values from 10^{-2} to 10^2. Instead of using all values of C in this interval, the step was increased 10 times every 10 measures

TABLE IV
LIBSVM LIBRARY KERNEL TYPES.

Kernel type	Expression
Linear	$k(x, x') = x^T x'$
Polynomial	$k(x, x') = (\gamma x^T x' + r)^d, \gamma > 0$
Radial basis function (RBF)	$k(x, x') = \exp(-\gamma \|x - x'\|^2), \gamma > 0$
Sigmoid	$k(x, x') = \tanh(-\gamma x^T x' + r)$

(logarithmic scale). This led to a total of 36 classifiers trained.

2) *Polynomial kernel*. The available parameters are C and degree (d). C took values from 10^{-3} to 10 (with 10 times increase in step) and d took values from 2 to 6 (with unit step increase). This led to a total of 25 classifiers trained.
3) *Radial basis function kernel (RBF)*. The available parameters are C and width (γ). C took values from 10^{-2} to 10^4 and γ took values from 10^{-5} to 10. In both axis the step was increased 10 times. This led to a total of 48 classifiers trained.
4) *Sigmoid kernel*. As the parameters are the same, the approach used for RBF kernel was repeated. This led to a total of 48 classifiers trained.

The parameters were previously loaded into the database and then used to train the classifiers. These were trained firstly for features of type 0 (only text, equal weights). Then the same procedure was carried out for all types of features, leading to a total of 1413 classifiers trained.

E. Classifiers evaluation

Regarding a particular category, web pages are classified as either belonging (positive instances) or not belonging (negative instances) to that category. The prediction made by the classifier can be described in a confusion matrix such as the one shown in Table V.

TABLE V
CONFUSION MATRIX.

	Predicted positive	Predicted negative
Actual positive	TP	FN
Actual negative	FP	TN

In the context of the classification of web pages within a particular category, the confusion matrix has four entries:

1) *True positives* (TP) are the number of web pages correctly classified as belonging to the category;
2) *False positives* (FP) refer to the number of web pages incorrectly classified as belonging to the category;
3) *True negatives* (TN) correspond to the number of web pages correctly classified as not belonging to the category;
4) *False negatives* (FN) refer to the number of web pages incorrectly classified as not belonging to the category.

Three evaluation measures can be defined from these entries:

1) *Precision* is the number of web pages correctly classified as belonging to the category divided by the total number of web pages predicted for that category (4);
2) *Recall* is the number of web pages correctly classified as belonging to the category divided by the total number of web pages that actually belong to that category (5);
3) *Accuracy* is the number of correctly predicted web pages divided by the total number of all web pages (6).

$$P = \frac{TP}{TP + FP} \quad (4) \qquad R = \frac{TP}{TP + FN} \quad (5)$$

$$A = \frac{TP + TN}{TP + FN + FP + +TN} \quad (6)$$

It is worth noting that the higher the precision the smaller the amount of misclassified web pages and higher recall means a smaller amount of missed correct web pages [31].

The *F1 measure* combines both precision and recall with an equal weight into a single parameter [32]. It is defined as follows:

$$F1 = \frac{2PR}{P + R} \quad (7)$$

The global estimates of a classifier performance are calculated using both macro-averaging (results are calculated on a per-category basis, then averaged over categories) and micro-averaging (results are calculated based on global sums over all classifications) of precision, recall and F1 measure [1]. As the F1 measure is a single parameter, it is used to compare performance of classifiers. N is the total number of categories (in this study N=6).

$$F1_{macro} = \frac{\sum_{k=1}^{N} F1_i}{N} \quad (8) \qquad F1_\mu = \frac{2P_\mu R_\mu}{P_\mu + R_\mu} \quad (9)$$

$$P_\mu = \frac{\sum_{k=1}^{N} TP_i}{\sum_{k=1}^{N} TP_i + \sum_{k=1}^{N} FP_i} \quad (10)$$

$$R_\mu = \frac{\sum_{k=1}^{N} TP_i}{\sum_{k=1}^{N} TP_i + \sum_{k=1}^{N} FN_i} \quad (11)$$

Micro-averaged measures tend to be dominated by the most commonly used categories, while macro-averaged measures tend to be dominated by the performance in rarely used categories [33]. According to the web pages distribution presented in Figure 2, there are three dominating categories (course, faculty and student) leading to 82% of the data set. For this reason, in this study the performance comparison was based on the micro-averaged F1 measure ($F1_\mu$).

All classifiers were tested using the testing data of the type of features they were trained with. In each evaluation, the confusion matrix for each category was calculated. This led to values for precision, recall and F1 measure per category. Then the overall evaluation measures were collected and $F1_\mu$ was calculated for each classifier.

V. Results and Analysis

As the purpose of this study is to measure potential improvement in web page classification by the inclusion of mark-up features, the type of features is the prime parameter for comparing results of each classifier. The SVM kernel type is also an important parameter in the training of classifiers, therefore, as four different types of kernel were used, the results were separated for each. Next subsection presents the classifiers performance per features type in each kernel type, varying the relevant parameters. This analysis is summarized by grouping all the best classifiers for each combination of features type and kernel type. Thus, it becomes evident which are the classifiers where web page classification was improved by including mark-up features. Finally, the best classifier overall is identified and its performance measures are presented in more detail.

A. Classifiers performance

1) *Linear kernel*. Figure 3 presents the classifiers performance across values of C for all types of features.

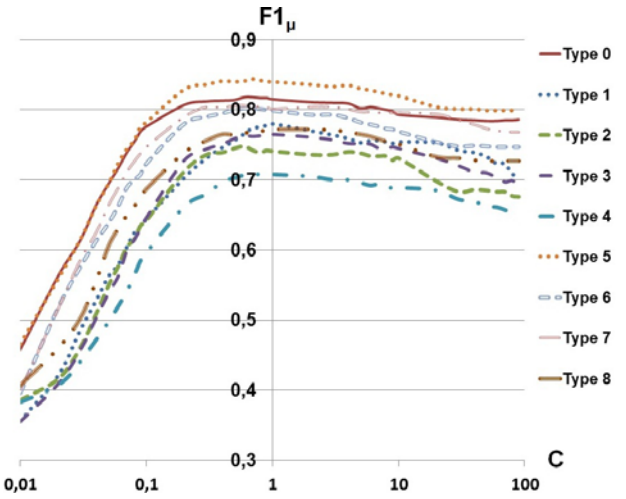


Fig. 3. Performance of linear kernel classifiers for all types of features.

From Figure 3 is it clear that classifiers trained with features of type 5 (text combined with different weighting in elements `TITLE` and `H1` to `H4`) have better performance than classifiers trained with features of other types. The value of $C = 0,7$ leads to the maximum performance with $F1_\mu = 0,818$ for features of type 0 (baseline) and $F1_\mu = 0,844$ for features of type 5.

2) *Polynomial kernel*. The polynomial kernel classifiers were insensitive to variations in degree (d) and C. However, variations per type of features were observed which are presented in Figure 4. Although all values of $F1_\mu$ are much lower than the ones observed for the linear kernel classifiers, it is noticeable a different behaviour in improvement per type of features.

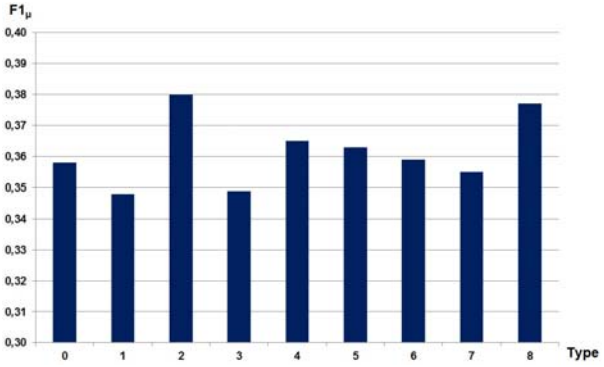


Fig. 4. Performance of polynomial kernel classifiers for all types of features.

Classifiers with polynomial kernel, trained with features of type 2 (only text in elements `TITLE`, `H1` to `H4`and `A`) outperform all trained with features of other types.

3) *RBF kernel*. The approach for comparison is a set of hit-maps with combinations of values of C and γ for each type of features. The results are depicted in Figures 5 and 6 for features of types 0 and 5 respectively. The values of $F1_\mu$ are color-coded according to the accompanying scale.

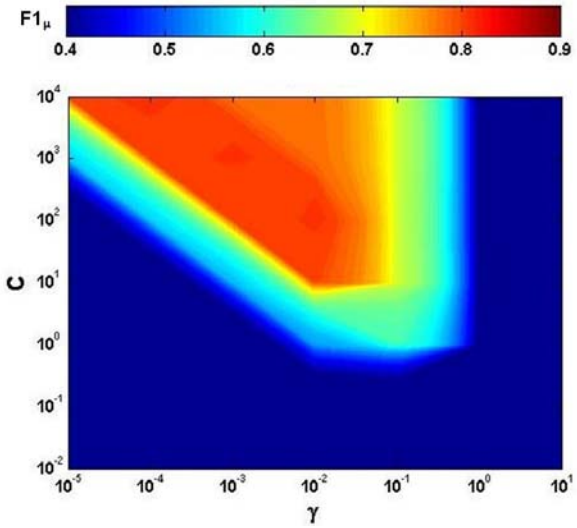


Fig. 5. Performance of RBF kernel classifiers for features of type 0.

The dark red areas represent combinations of C and γ that lead to $F1_\mu$ higher than 0,8. The maximum result observed in a RBF kernel classifier was $F1_\mu = 0,834$ for features of type 5 (with $C = 10^4$ and $\gamma = 10^{-4}$).
For comparison purposes, Figure 7 shows a summary of the performance of classifiers trained with features of types 1 to 8. Besides features of type 5, only classifiers trained with features of type 7 show a dark red area (much smaller in size), but with $F1_\mu = 0,806$ (with $C = 10^3$ and $\gamma = 10^{-3}$).

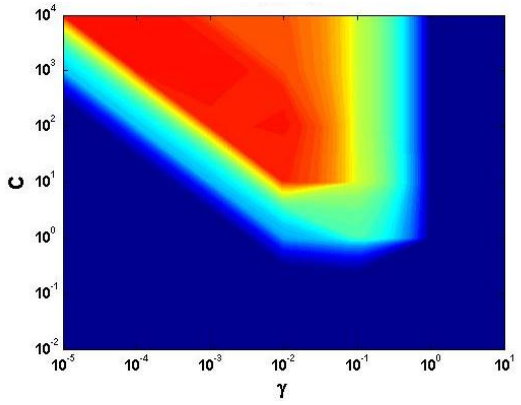


Fig. 6. Performance of RBF kernel classifiers for features of type 5.

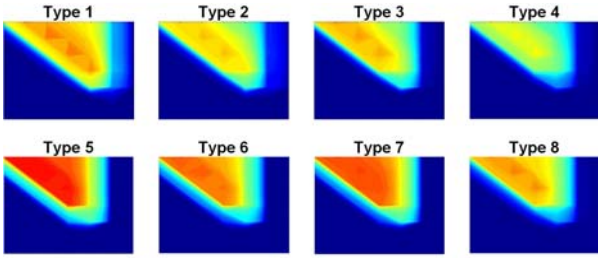


Fig. 7. Performance of RBF kernel classifier for features of types 1-8.

4) *Sigmoid kernel*. The approach for comparison is again a set of hit-maps as the parameters are the same as used for RBF kernel classifiers. Results are depicted in Figure 8 for features of types 0 and 5.

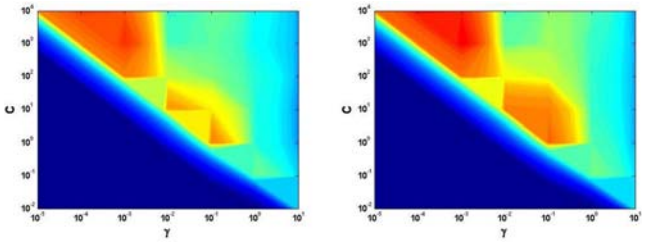


Fig. 8. Performance of sigmoid kernel classifiers for features types 0 and 5.

The maximum result observed for sigmoid kernel classifiers was $F1_\mu = 0,840$ for features of type 5 (with $C = 10^3$ and $\gamma = 10^{-3}$). Figure 9 shows the same analysis for feature of types 1 to 8. The same behaviour noticed for RBF kernel classifiers is also evident in sigmoid kernel classifiers: besides features of type 5, only features of type 7 show some improvement. This confirms the results for the linear kernel and RBF classifiers.

B. Comparison of improvements in classification performance

Improvements in web page classification can be identified by comparing the values of $F1_\mu$ in each classifier. Table VI

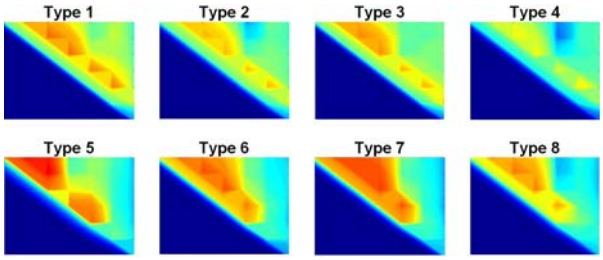


Fig. 9. Performance of sigmoid kernel classifiers for features of types 1-8.

presents the classifiers with best performance for each combination of features type and kernel type (with its corresponding parameters).

TABLE VI
BEST CLASSIFIERS FOR EACH FEATURE AND KERNEL TYPES.

Feature type	Linear $F1_\mu$	C	Polyn. $F1_\mu$	RBF $F1_\mu$	C, γ	Sigmoid $F1_\mu$	C, γ
0	0,818	0,7	0,358	0,813	10^2,10^{-2}	0,815	10^3,10^{-3}
1	0,780	1,0	0,348	0,777	10^3,10^{-3}	0,780	10^2,10^{-2}
2	0,748	0,6	**0,380**	0,742	10^3,10^{-3}	0,745	10^0,10^0
3	0,765	1,0	0,349	0,762	10^3,10^{-4}	0,765	10^2,10^{-2}
4	0,709	0,9	0,365	0,710	10^2,10^{-2}	0,722	10^0,10^0
5	**0,844**	0,7	0,363	**0,834**	10^4,10^{-4}	**0,840**	10^3,10^{-3}
6	0,802	0,7	0,359	0,796	10^3,10^{-3}	0,800	10^2,10^{-2}
7	0,806	0,6	0,355	0,806	10^3,10^{-3}	0,806	10^2,10^{-2}
8	0,772	1,0	0,377	0,773	10^3,10^{-3}	0,772	10^3,10^{-3}

From Table VI it is possible to conclude that classifiers trained with features of type 5 are the only ones that show improvement in performance (relatively to baseline which are classifiers trained with features of type 0). This shows that:

1) using only mark-up specific features is not enough, as it actually decreases performance (see results for classifiers trained for features of types 1, 2, 3 and 4).
2) even combined with text, as more mark-up specific features are added, performance also decreases (see results for classifiers trained for features of types 6, 7 and 8).

This means that improvement in performance arises from the combination of features associated with specific HTML elements with features coming from the rest of the text. Words in elements such as title and headers clearly seem to have more discriminative power than words in other elements and help improving the quality of web page classification. This is observed in linear, RBF and sigmoid kernel classifiers.

The only exception to this behaviour is for polynomial kernel classifiers where improvements are revealed for features of types 2 and 8. As these classifiers have a much lower performance than all others, they are considered not relevant for this analysis. Thus, their behaviour is not taken into account for the overall conclusion.

C. Best classifier overall

Table VI also shows which is the best classifier overall. It is a linear kernel with $C = 0,7$ and trained for features of

type 5 (text and different weights given to words in `TITLE`, `H1`, `H2`, `H3` and `H4` elements) with $F1_{\mu} = 0,844$.

The confusion matrix for all categories is presented in Table VII.

TABLE VII
BEST CLASSIFIER ALL CATEGORIES CONFUSION MATRIX.

Actual categories	Predicted categories course	department	faculty	project	staff	student
course	272	0	1	3	0	11
depart.	1	37	3	0	0	6
faculty	4	1	246	4	1	63
project	4	1	7	121	0	31
staff	2	0	12	0	1	31
student	7	0	17	2	0	466

The overall and per category performance is presented in Table VIII.

TABLE VIII
BEST CLASSIFIER PERFORMANCE.

Categories	Precision	Recall	Accuracy	F1
course	0,938	0,948	0,976	0,943
department	0,949	0,787	0,991	0,860
faculty	0,860	0,771	0,917	0,813
project	0,931	0,738	0,962	0,823
staff	0,500	0,022	0,966	0,042
student	0,766	0,947	0,876	0,847
Macro-averaged	0,824	0,702	0,948	0,721
Micro-averaged	0,844	0,844	0,948	0,844

VI. CONCLUSION

Experimental results show that including structural information from mark-up can improve performance of web page classification. However, this effect only happens in a particular set of circumstances. Improvement is observed when combining a term frequency analysis with differently weighting of terms occurring in the title and header HTML elements of a web page. The same approach applied for terms extracted from other structural information (text enhancement, anchors or list items) decreases the quality of classification. This means that adding more features only is not the solution to improve a classifier, even if they come from the structure of the web page.

In this study, the observed improvement is not very significant (around 3%) and has no big impact in performance. Yet it points out that key feature selection techniques is the path to improve the performance of a classifier. Although there is evidence that including mark-up features benefits a classifier performance, it is also apparent that these have to be chosen carefully. This prompts that future work should focus, particularly, in the combination of mark-up specific features with feature reduction techniques in order to achieve greater impact in improving web page classification performance.

REFERENCES

[1] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*.

[2] A. Kilgarriff and G. Grefenstette, "Introduction to the special issue on the web as corpus," *Computational linguistics*, 2003.

[3] "WebCorp: The Web as Corpus." [Online]. Available: http://www.webcorp.org.uk

[4] F. Sebastiani, "Text Categorization," no. Ml.

[5] X. Qi and B. Davison, "Web page classification: Features and algorithms," *ACM Computing Surveys (CSUR)*, no. June, pp. 1–31, 2009.

[6] D. Mladenic, "Turning Yahoo into an Automatic Web-Page Classifier," 1998.

[7] K. Golub and A. Ardö, "Importance of HTML Structural Elements and Metadata in Automated Subject Classification," pp. 368–378, 2005.

[8] "ODP - Open Directory Project." [Online]. Available: http://www.dmoz.org/

[9] C. Chekuri and M. Goldwasser, "Web search using automatic classification," *... on the World Wide Web*, 1997.

[10] M. Sinka and D. Corne, "A large benchmark dataset for web document clustering," *Soft Computing Systems: Design, ...*, 2002.

[11] X. Qi and B. D. Davison, "Knowing a web page by the company it keeps," *Proceedings of the 15th ACM international conference on Information and knowledge management - CIKM '06*, p. 228, 2006.

[12] B. Liu, X. Li, W. Lee, and P. Yu, "Text classification by labeling words," *Proceedings of the National Conference on ...*, 2004.

[13] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," *KDD workshop on text mining*, pp. 1–20, 2000.

[14] S. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Frontiers in Artificial ...*, vol. 31, pp. 249–268, 2007.

[15] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Machine learning: ECML-98*, 1998.

[16] D. Riboni, "Feature selection for web page classification," *EURASIA-ICT 2002 Proceedings of the Workshop*, 2002.

[17] S. Shibu, A. Vishwakarma, and N. Bhargava, "A combination approach for Web Page Classification using Page Rank and Feature Selection Technique," *International Journal of Computer ...*, vol. 2, no. 6, 2010.

[18] S. Ozel, "A genetic algorithm based optimal feature selection for web page classification," *Innovations in Intelligent Systems and Applications (...*, pp. 282–286, 2011.

[19] Selvakuberan, Indradevi, and Rajaram, "Combined Feature Selection and classificationA novel approach for the categorization of web pages," *Journal of Information ...*, vol. 3, no. 2, pp. 83–89, 2008.

[20] J. Alamelu Mangai, V. Santhosh Kumar, and S. Appavu alias Balamurugan, "A novel feature selection framework for automatic web page classification," *International Journal of Automation and Computing*, vol. 9, no. 4, pp. 442–448, Aug. 2012.

[21] "The 4 Universities Data Set." [Online]. Available: http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

[22] M. Craven, A. McCallum, and D. PiPasquo, "Learning to extract symbolic knowledge from the World Wide Web," 1998.

[23] N. Ali and N. Ibrahim, "Porter Stemming Algorithm for Semantic Checking," pp. 253–258, 2012.

[24] J. Abernathy, "PHP Class: Porter Stemming Algorithm." [Online]. Available: http://www.chuggnutt.com/stemmer-source

[25] "English Stopwords." [Online]. Available: http://www.ranks.nl/resources/stopwords.html

[26] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[27] K. Senathipathi and K. Batri, "Keystroke Dynamics Based Human Authentication System using Genetic Algorithm," *European ...*, vol. 82, no. 3, pp. 446–459, 2012.

[28] P. Juszczak, D. Tax, and R. Duin, "Feature scaling in support vector data description," *Proc. 8th Annual Conf. of the Advanced ...*, 2002.

[29] C.-C. Chang and C.-J. Lin, "LIBSVM - A Library for Support Vector Machines." [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[30] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," *Proceedings of the 23rd international conference ...*, pp. 233–240, 2006.

[31] J. Pierre, "On the automated classification of web sites," *arXiv preprint cs/0102002*, vol. 6, no. 0, 2001.

[32] Y. Yang and X. Liu, "A re-examination of text categorization methods," *Proceedings of the 22nd annual international ACM ...*, pp. 42–49, 1999.

[33] A. Özgür and T. Güngör, "Classification of skewed and homogenous document corpora with class-based and corpus-based keywords," *KI 2006: Advances in Artificial Intelligence*, 2007.

Towards a virtual population of drivers: using real drivers to elicit behaviour

Joel Gonçalves
LIACC
Faculty of Engineering
University of Porto
Porto, Portugal 4200-465
Email: pro12009@fe.up.pt

***Abstract*—Driving behaviour plays a fundamental role in transportation systems, where each single driver has a unique behaviour. In this paper, we propose a methodology based on low-cost simulators for collecting data, we use the dynamic time warping technique for assess the similarity between behaviours, and hierarchical clustering for grouping drivers with similar behaviours. With this approach we expect to extract driving behaviour at individual level, form groups and characterize them in fast, cheap and methodological way. The preliminary results have achieved the desired goals, where we were able to identify multiple clusters for longitudinal and lateral control metrics with a small sample of drivers.**

I. Introduction

Nowadays, Transportation System research is concerned with several important issues related to rapid development of traffic network, specially in urban areas. This phenomena has introduced dramatic changes in citizens mobility and quality of life. Furthermore, it has proved to be a difficult challenge to cope with by researchers, decision-making, and practitioners [1]. While an adequate transportation system enables a good experience for the users, the contrary may be responsible for economic, social, and environmental issues.

By its nature, a transport system can easily become far too complex to be modelled with traditional mathematical approaches [2]. The elements composing the system (e.g: pedestrians, vehicles, road network layout, signalling layout, control systems, and so forth...), the interactions between, and the solution space for solving a specific problem can be overwhelming. Under such conditions, simulation emerges as a natural approach for handling this complexity. Using simulation, one can model the desired transport system, explore applicable actions to the system, and predict the overcoming results for each action [3,4]. This approach gives the advantage of covering a vast space of solutions in short time and without disrupting the real system. However, we cannot ignore the significant challenges for modelling the system in a simulation, specially when the problem to be solved may be influenced by multiple entities with their specific interactions and dependencies[2]. For example, reducing the traffic congestion on a road intersection using traffic light's times we should take into account the number of vehicles and pedestrians using the intersection; to increase the complexity of modelling such system, traffic lights could change their behaviour to help other intersections, the pedestrians may change their habits according to the new traffic lights' behaviour, and regular drivers may also react to the changes; another possibility would be to change the traffic network itself around the intersection with the associated cost, time, disruption and expectation on how the elements would react during the construction and after the implementation of the changes. As the example tries to demonstrate, the elements composed in the system are all interconnected, they react to each others but always ensuring their autonomy to pursue their objectives. For the rest of the paper, we will focus on traffic simulation.

The use of Multi-Agent Systems (MASs) as a paradigm for modelling the Transportation Systems rapidly emerged [5-7]. Specifically, micro traffic simulators have the ability to represent each individual vehicle in the transportation system. Each of these vehicles represent a driver, with a pre-defined starting point and destination point. Depending on the network, the drivers may choose their own path according to some decision-making process, they also may change lanes to take over other slower vehicles. Albeit these simulators give a coherent solution for analysing some problems, most of the criticism to this approach is focused on the validation of these tools. In this paper, we intent to contribute with a methodology for evolving the rigid and predictable traffic simulator vehicle behaviours with behaviours that mimic real drivers' behaviour. Our final hypothesis is that if in our simulations we create a virtual population of drivers where each of them resembles a set of extracted driving behaviour patterns, then our simulations will inherit driving behaviour validation and our predictions will be more accurate than traditional driving behaviour approaches. Unfortunately, a long path is yet to be pursued to achieve that goal but in this paper we will present our current state on driving behaviour elicitation.

Concerning the structure of this paper, in chapter II we introduce two major tools used for the behaviour elicitation process, then in chapter III we present our envisioned methodology. After that, chapter IV reports an experiment performed, we show the results obtained in chapter V, and in chapter VI we review some related work relevant for this topic. Finally, in chapter VII we give an overview of this specific work and project future directions.

II. Background

A. Low-Cost Driving Simulators

Driving simulators have been used for developing research [8,9]. In this type of simulation, we focus on how drivers

control the vehicle along the scenario. Specifically, their lateral position management in the lane and their longitudinal control behaviour though velocity and headway distance to a reference vehicle. In this context, simulation is used for overcoming the dangers to the participants, the vehicle maintenance cost and possibility to design a scenario to address specific needs.

Albeit driving simulators and microscopic traffic simulators operate at individual level, the former distinguish by focusing the experience in the human interaction with the controlled vehicle and the vehicles around as the driver progresses through the scenario. This interactivity implies very dynamic simulation results which are possible to observe by taking several drivers to drive the exact same vehicle, and the exact same road structure. In such conditions, it is expected to have significant longitudinal and lateral control behaviours. For this work we will focus on the use of low-cost driving simulators. The most interesting characteristics [10] are:

- **Cost.** This specific type of simulator, has opposed to high fidelity simulators has a very reduced cost due to the use of commercial hardware and software.
- **Reproducibility.** The same software and hardware setup can be easily obtained. Furthermore, since the simulation environment is controlled we can reproduce the same scenarios and situation and expose them to new subjects.
- **Dissemination.** Also related with the reproducibility, this feature allows the dissemination of a simulation setup across distinct geographically distributed locations. We can take advantage in terms of easing the subjects access to the simulator.
- **Fidelity.** Albeit not as reliable as high fidelity simulators, there are several studies that show evidence of significant correlations between low-cost driving simulators and real vehicles.

In this work we will use low-cost simulators for collecting the driving performance data.

B. Dynamic Time Warping

The Dynamic Time Warping (DTW) is a class of algorithms design to compute a distance measure between two time series which may have distinct lengths [11].

A traditional euclidean distance could be computed if the time series have the exact same length and, more important, their frequency of events is synchronized. However, in our problem domain, a typical example would be the velocity time series of two distinct drivers in a circuit. Since the simulator is continuously logging the velocity, at fixed instants, it is intuitive to predict that the slower driver would have more log instances than the faster one. This would bring the following problems to calculate a distance measure:

- **Length.** A traditional euclidean distance would require equal length series to calculate the distance between points at the same index.
- **Alignment.** Even in the advent of equal length series, a common phenomena is misalignment between series. In practice, this means that despite we can understand visually that two series have a similar pattern, the euclidean approach would be too sensible to the misalignment and fail to capture the similarity between the series.

Formalizing, the DTW technique receives as input two series: Q and R. The Q series is defined as $X = (x_1, ..., x_N)$, and R as $Y = (y_1, ..., y_M)$. For simplicity we define $i = 1...N$ while $j = 1...M$ for the indexes of Q and R, respectively. Also, we define a non-negative $localdistance$ function f that calculates the distance between the points x_i and y_j, as defined with:

$$d(i, j) = f(x_i, y_j) \geq 0 \quad (1)$$

After calculate each local distance, the algorithm proceeds by computing the $warping\ path$, defined as $W = w_1, ..., w_K$ where each $w_k = (i, j)$ such that:

$$MAX(m, n) \leq K \leq m + n - 1 \quad (2)$$

K is defined as positive number corresponding to the warping path index. Meanwhile, this path must hold the following conditions:

- **Boundary:** $w_k = (1, 1)$ and $w_K = (N, M)$.
- **Monotony:** $x_1 \leq x_k \leq x_K$ and $y_1 \leq y_k \leq y_K$.
- **Step size:** $w_{k+1} - w_k \in \{(1, 0), (0, 1), (1, 1)\}, k \in [1 : K - 1]$.

The boundary condition defines that the $warping\ path$ must start in the first points and end in the last points of each series. As for the monotony it ensures that the algorithm converges to the upper edge of the matrix. Finally, the step size condition forces the path to travel only to adjacent cells.

The $global\ distance$ between series is:

$$D(X, Y) = \sum_{k}^{K} d(x_{n_k}, y_{m_k}). \quad (3)$$

Where the distance between both series is the sum of the local distances corresponding to the warping path.

The DTW technique overcomes length and alignment mismatchings and is able to deliver a global distance metric that defines how similar the two series are. This technique can prove useful for building distance matrix between the different series which is a convenient input for many clustering techniques.

III. Methodology

Overall our methodology is decomposed in three sequential steps: (i) driver behaviour modelling, (ii) behaviour elicitation, and (iii) population generation and validation. We describe the steps in the following sub-chapters.

A. Driver Behaviour Modelling

The importance of this step resides in the formulation of the driving behaviour and mapping between driving performance measurements and driver behaviour subtasks [12,13].

A driver behaviour module is a design structure that identifies and isolates driving subtasks into modules and describes

how those modules interact between themselves. Multiple driver behaviour models can be found in the literature. For the purpose of this work we will focus on the Extended Control Model (ECOM), proposed in [13]. This model has the advantage of being very adaptable to the driving task, their modules have smaller granularity than most models, and they can be easily extended to other type of subtasks. The ECOM is composed by the tracking, regulating, monitoring and targeting layers. The layers assume a hierarchical configuration where information flows in a loop from the top to the bottom, although it is admitted that some interactions may reverse the information flow. At top we have the targeting layer to manage the higher goals of our travel, in other words defines our destination and our travel path. Then, the monitoring layer handles the vehicle/environment observation, which extracts information concerning spatial position of the vehicle and signals relevant to the driving task. After that, the regulating layer defines the desired driving performance state to the specific context. At last, the tracking layer ensures that the driving performance state defined above are kept at satisfactory level.

For the sake of simplicity we will focus on the regulating layer. The targeting layer in most simulations are quite static (vehicle have their destination and path pre-defined), while the monitoring and tracking layers are highly dependent on the input/output interfaces available in the simulators. Hence, we focus on the regulating layer, which give us abstraction from simulation dependent features and focus on how drivers adapt to specific situations. We define the input from the monitoring layer as $M = (Hv, Rv, G, L, S)$, where Hv is the host vehicle, the Rv is the reference vehicle, G is information associated to the gap between the host and reference vehicle, L is the road layout (e.g.: curve or straight line), and S traffic signals applicable in the current context. Both Hv and Rv are multi attribute variables defining the respective vehicles, such has positioning, orientation and velocity. This formulation simpler form would be the absence of other vehicles, where the Rv and G would be inexistent. As other vehicles share the space environment interaction take place between the host and the reference vehicle. For the purpose of this paper we will focus on the single vehicle scenario, where the road layout will assume the the major relevance. For this situation we defined the following output from the regulating layer as $R = (P_{des}, P_{tol}, V_{des}, V_{tol})$. The P_{des} and P_{tol} are the desired position within the lane and the tolerance margin drivers willing to accept, respectively. The same meaning applies to V_{des} and V_{tol} but in terms of velocity. As experiments reveal new information, the model used to structure the driver behaviour may suffer changes, specially in terms of relevant variables for each layer.

B. Behaviour Elicitation

Behaviour elicitation, in this context, is the act of extracting the driver behaviour from a sample of real drivers by monitoring their driving performance in a specific driving context. Since we aim to simulate entire population of drivers, we identified the need to have tools that ease in this process of collecting data and in the absence of all real drivers, we should be able to generate a representative population based on the available driving behaviours. For addressing the data collection task, we will use low-cost driving simulators. The advantages of using these tools are enumerated in II-A. As for the generation of population based on a sample, we aim to perform clustering over the collected driving behaviour in order to obtain representative groups that can be used to generate similar behaviours.

C. Population Generation and Validation

The population generation task is highly dependent on the simulator used. Nevertheless, it seems the agents must have at least the following basic characteristics: satisfier approach and a low-resource consumption. Concerning the driving behaviour, agents should tend to act as satisfiers, meaning that they have preference values according to the output from the regulating layer but agents would only adjust their driving performance if they are outside a "comfort zone". Also very important, due to the dimension of the population is that each individual agent should consume low memory and processing time in order to the simulation handle a large population at the same time. In respect to validation we propose a two step process: (i) for validate the elicitation process, we will select a set of drivers, extract their driving behaviour, generate the respective agents, and compare the driving performance of both agents and real drivers in a new map; (ii) at population level, assuming the previous validation step was successful, use a high dimension sample of real drivers, generate the population and compare the results between real-system, the agent based simulator and a third high fidelity simulator.

IV. Experiment

In this paper, we will address the behaviour elicitation aspect of the methodology, more specifically we intent to capture the driving behaviour on a simple scenario where drivers are the only mobile entities on the map. Once the data is collected we can then apply data mining techniques for extracting useful clusters and statistics.

For the sake of clarity, this is part of a work in progress, hence important features of the methodology are not yet addressed. For example, a ECOM driver model adaptation for artificial intelligence entities, the agent model training based on the elicited behaviour, and the process validation are out of the scope of this paper.

A. Design

Experiments with volunteers were conducted to collect data. The experiments begun with a briefing to make sure people are calm, understand the objective of the study, they should respect the velocity limit of 120 Km/h and the pavement markings.

After the initial step, the volunteers were encouraged to experiment the pedals and the steer to adapt to the respective feedback. They were able to train as much time as they desire along the same map of the experiment without recording the data for experiment purposes. When they felt ready, the actual experimental data was logged. Each driver had to complete 3 laps around the map. Upon completing one lap, the game was terminated by the experiment supervisor, the volunteer would rest for 1 minute, then proceed to the next lap. Each lap took between 3-5 minutes depending on the drivers velocity.

We achieve a total of 9 participants, all male gender. From this sample, none had use the simulator before. However, User 6 (henceforth we will address each volunteer as User) reported to be the only one with previous experience with the steering wheel and pedals equipment. As for the map, we will use a simple rural map with the format of round corner square, as illustrated in Fig 1. The road is composed by two lanes with opposite directions. There where no other vehicles in the map either in same or opposite direction.

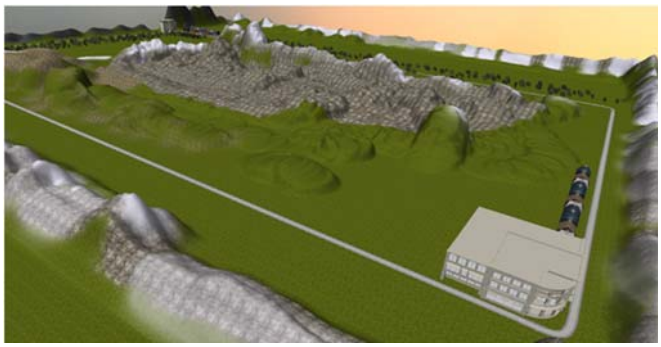


Fig. 1: Snapshot of the map. It has the shape of a rectangle with round corners.

B. Setup

The setup of our experiment was installed at LIACC's lab. The hardware features were a 40" HD 1080p television, a desktop computer for running the simulation software, and a Logitech G27 Racing Wheel with steering wheel and pedals. In this experiment, the vehicle was set with automatic gear change, hence both gear pedal and shifter were deactivated. The sound was delivered by the TV's built-in speakers.

Fig. 2: A user taking the experiment in the low-cost driving simulator.

In terms of simulation software we used the Serious Driving serious game for implementing the desired experiments. Users had a similar perspective of the surrounding driving environment, from a position inside the cockpit.

The feedback provided by the simulation was mainly the visual translation of the vehicle throughout the map, but it also delivered a speedometer, the simulation's steering wheel was animated according to the real steering wheel's rotation, and the real steering wheel was able to provide haptic feedback using two motors built-in. A picture of the setup is available in figure 2.

C. Variables

The variables captured in the simulation were logged into a CSV file. In the end of the experiment we had the data with the following format:

- **Timestamp.** The clock time when the data was recorded.
- **Vehicle Position.** The vehicles global position in the map. Define by real numbers in the X and Z axis.
- **Vehicle Velocity.** The vehicles instant velocity for the X and Z axis. Defined by 2 vectors: X and Z velocity.
- **Pedals.** The force applied by the users on the acceleration and brake pedals. Both values range between 0 if none to 1 if the pedal is totally pressed.
- **Steer Angle.** The steering wheel angle. The value range from 0 for full right, 90 to normal position and 180 degrees for full left.
- **Right Wheel Position.** The position in terms of X, Z of the right wheel.

Since the simulation runs around 60 fps, the number of lines each file contains are around 11000 and 17000, which is the equivalent to approximately 3,3 minutes and 4,7 minutes respectively.

D. Post-processing

For achieving the desired values for longitudinal and lateral control, a new data set was computed before applying the algorithms. The post-processing phase can be divided in the following steps:

- **Series merge.** Combine the 3 samples from each User in a single series.
- **Velocity computation.** Compute the instant velocity using both velocity vectors.
- **Lateral position computation.** Compute the distance to the right side of the lane.
- **Map segmentation.** Segment the data, in terms of straight lines and curves.

During driving simulated experiments, users tend to perform higher number of driving errors. Several conditions contribute for this problem: anxiety for feeling evaluated, the awkwardness of simulated driving compared to real, sensation of safety in case of performing bold manoeuvres and the adaptation to the vehicle dynamics. This higher number of errors is not considered representative of the drivers driving behaviour, e.g. half of the curves, the driver went a little off road. We took 2 measures to minimize this problem: user training period

before the experiment and series merge. The later solution is performed by taking the smallest of the 3 series, and computes the average to the correspondent point from the other 2 series in the same index. With this procedure we assume that if drivers upon performing the same manoeuvre repeatedly, then they tend to have similar driving performance. In practice, this assumes that multiple series from the same driver will have approximated size and wave patterns. After doing the merge procedure, the new series tends to be "smother" and compensates local maxima/minima from the previous series. Furthermore, to minimize the weight of data size in computation, we reduced each series size in half by creating new series with only the values from the even indexes (giving a 30 points per second).

The instant velocity computation was performed using equation 4. Then, the instant velocity was converted from m/s to Km/h.

$$v = \sqrt{{v_x}^2 + {v_z}^2} \tag{4}$$

Concerning the lateral position computation, a script was written in Java language to handle that task. The function receives the coordinates from the vehicle's front right wheel and measures the distance to the right side of the lane. The map segmentation is explained in section IV-E.

E. Exploratory Data Analysis

An exploratory data analysis was conducted to have a global perspective of the data obtained after the post-processing step.

TABLE I: Descriptive statistics from a driver sample.

	max	mean	median	min	mode	std	var
x	1499,05	1023,07	1078,55	500,01	1497,33	440,13	193710
z	1899,39	942,55	882,97	99,16	167,34	717,58	514923,1
vel	124,27	92,56	106,03	0	0	27,72	768,35
steer	165,09	94,65	90,70	70,80	90,35	12,03	144,60
acc	0,55	0,12	0,12	0	0	0,06	0,003
brk	0	-0,02	0	-0,35	0	0,06	0,003
lp	16,24	1,77	2,09	-12,39	0,98	1,86	3,46

The X and Z variables denote the position along the map. Velocity limits (defined to 120 Km/h) seem to be fairly respected with the maximum of 124 Km/h and average of 92Km/h with 27,7 Km/m of standard deviation. As an important metric, lateral position acquisition it was disappointing to check that the metric is not being correctly measured from the simulator. Both minimum and maximum values are too extreme which evidences the existence of measurement errors, since these specific values were not observed during the experiments. Albeit this problem makes the stereotype task unreliable, we still consider the metric useful for clustering since the series pattern is somehow similar in all data.

Map segmentation task was performed by classifying segments of the map as straight line or curve. There are significant behaviour changes when drivers are driving a straight line or a curve. In figure 3 there is a sample data taken from a user, which presents the velocity around the map. It is clear the user systematically, when approaching a curve, starts decreasing the velocity until a minima (around the middle of the curve). Then as the user leave the curve its velocity starts to increase to levels around previously enter the curve. When in straight line, the velocity tends to be stable between the 100 Km/h and 120 Km/h.

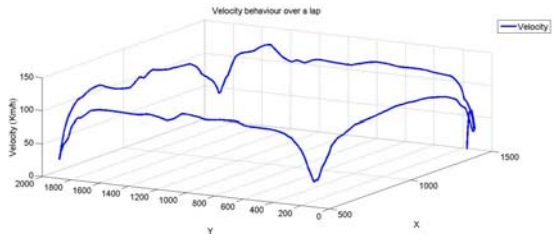


Fig. 3: A 3D plot illustrating the velocity behaviour as the vehicle moves through the map.

The line segments were defined as a line that starts 350 m after a curve and ends 350 m before a new curve. As for the curve, was defined as a curved segment which starts 300 m before and 300 m after the actual curve.

V. Results

A. Straight line behaviour

Straight line behaviour results showed the existence of 4 clusters for the velocity and 2 clusters for the lateral position along the lane. The cophenetic correlation coefficients (CCC) were 0,91 and 0,76 respectively. This metric assess the correlation between the distances calculated in the hierarchy tree and the observed data. We define a straight line cluster for the velocity as $S_{vel} = (S_{vel_1}, ..., S_{vel_P})$ and the straight line cluster for lateral position as $S_{lp} = (S_{lp_1}, ..., S_{lp_Q})$.

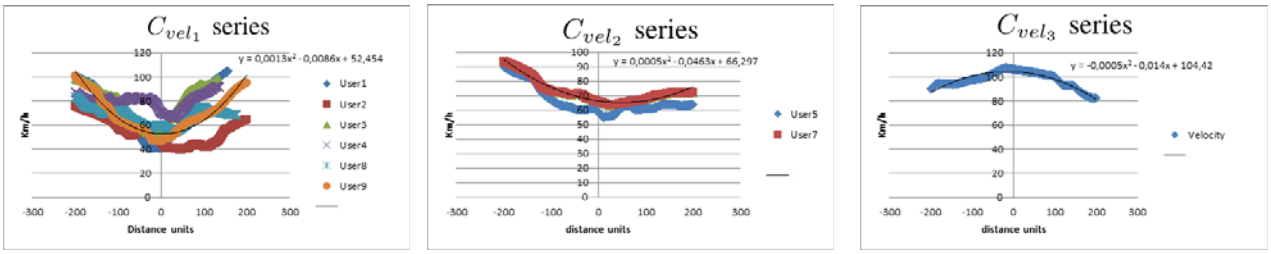
TABLE II: Descriptive statistics of each straight line velocity cluster. Units in Km/h.

Cluster	Max	Mean	Med	Min	Mode	STD	Var	Users
S_{vel_1}	118,19	108,50	110,46	90,38	109,77	6,08	36,93	U_1, U_5, U_9
S_{vel_2}	116,62	103,60	105,64	80,52	107,52	8,18	66,97	U_3, U_4, U_7, U_8
S_{vel_3}	113,90	99,98	100,49	85,33	100.39	6,99	48,81	U_6
S_{vel_4}	91,34	80,88	81,22	69,33	80,84	4,98	24,82	U_2

The velocity clusters population weren't very balanced. While S_{vel_1} and S_{vel_2} clusters have 3 and 4 elements respectively, S_{vel_3} and S_{vel_4} only had 1 element each. Significant differences between S_{vel_1} and S_{vel_2} where not found, however significant differences were observed between S_{vel_1} and S_{vel_3}, S_{vel_1} and S_{vel_4}, and S_{vel_3} and S_{vel_4}.

TABLE III: Descriptive statistics of each straight line lateral position cluster. Units in m.

Cluster	Max	Mean	Med	Min	Mode	STD	Var	Users
S_{lp_1}	2,44	2,11	2,12	1,64	2,16	0,14	0,02	U_2
S_{lp_2}	2,85	2,05	2,10	-2,21	2,95	0,39	0,15	$U_1, U_3, U_4, U_5, U_6, U_7, U_8, U_9$



(a) C_{vel_1} (b) C_{vel_2} (c) C_{vel_3}

Fig. 4: Graphical representation of velocity, as drivers handle a curve, by cluster.

Lateral position clusters highlight U_2 behaviour as different from the rest of the sample. However, no significant difference between the two statistics.

B. Curve behaviour

Our research identified 3 velocity behaviour clusters drivers engage to handle curves. For better perception of the clusters, the corresponding scatter plots are presented in figure 4.

TABLE IV: Descriptive statistics of each curve velocity cluster. Units in Km/h.

Cluster	Max	Mean	Med	Min	Mode	STD	Var	Users
C_{vel_1}	104,59	69,85	69,56	39,67	97,58	15,81	249,86	$U_1, U_2, U_3, U_4, U_8, U_9$
C_{vel_2}	93,29	69,61	67,46	55,28	71.25	9,06	82,05	U_5, U_7
C_{vel_3}	107,48	97,69	98,10	82,39	96,95	6,60	43,62	U_6

A proposed characterization of the clusters is available in table IV. The C_{vel_1} cluster had the majority elements assigned with two thirds of the overall sample, followed by C_{vel_2} with two elements and C_{vel_3} with only one. Statistically comparing the distributions we found significant differences between C_{vel_1} and C_{vel_3}, and C_{vel_2} and C_{vel_3}.

VI. Related Work

There are numerous works for solving transportation challenges using artificially intelligent agents [14]. In [15] identifies the opportunity for improving Advanced Traveller Information Systems (ATISs), the author states that drivers should be modelled as social agents with the ability to coordinate decisions. According to the ECOM driving model, such work could be useful to the targeting layer, however they do not considered the actual driving behaviour. Other works that may also help specify the drivers' decision process can be found in [16-17]. Works such as [18-21] work in a "tactical" level. This widely used term refers to tackle the modelling of driving subtasks addressed in the ECOM's regulating layer. Albeit their modelling effort is interesting, the model instantiation is more based on researchers experience than in actual drivers. In our opinion this somehow defeats the purpose of modelling complex behaviours since in the end the models are instantiated with subjective values such as the majority of simulators.

To our best knowledge, there are not ongoing research on eliciting human driving behaviour for creating driving agents.

VII. Conclusion

In this work we captured driving data using a low-cost serious game for extracting behaviour patterns. To accomplish such goal we use a data mining approach for tackle the problem.

After collecting the data from the simulator, a post-process phase was initiated to transform data to more suitable variables. Still in this phase, we used multiple samples to smooth data captured in order to minimize the effects of driving simulators that tend to produce bad data. Then, after analysing the initial data we decided to segment the map in zones of straight line and curves.

For calculating the similarity between time series we used the Dynamic Time Warping algorithm. This technique provides a distance measure value and it is capable to cope with time series not synchronized. Upon obtaining the distance matrices, we could then use aggregative hierarchical clustering for grouping users. Then, for each cluster we obtained the descriptive statistics and the trend lines when adequate.

With this methodology we were able to find 4 clusters for velocity in a straight line, and 3 for curves. As for the lateral position we obtained 2 clusters and 3 cluster for straight line and curves, respectively.

For future work, we need to solve the lateral position bug in the curves cases. So we can understand how drivers use their position to handle the curves. Furthermore, the number of volunteers must also raise to obtain more diverse data and (probably) obtain more clusters, or reinforce the ones already found. Effort must also be made to convert the driving behaviour into a lightweight agent for a driving simulator, and to study driving behaviour adaptation when multiple vehicles coexist in the same scenario.

Acknowledgment

The author would like to thank professor João Moreira and professor Carlos Soares for their advice and support in the improvement of this data mining based approach. Also, professor Rosaldo Rossetti for his supervision and feedback throughout all work.

I would also like to thank LIACC and FCT for proving the space and material for the experiments, respectively.

REFERENCES

[1] G. Dimitrakopoulos, *Intelligent Transportation Systems*. IEEE Vehicular Technology Magazine, vol. 5, issue 1, pp 77-84, 2010.

[2] Z. Kokkinogenis, L. Passos, R. Rossetti and J. Gabriel, *Towards the next-generation traffic simulation tools: a first evaluation*. Doctoral Symposium on Informatics Engineering , 2011.

[3] B.C. Silva, A. Bazzan, G.K. Andriotti, F. Lopes, D. Oliveira, *Itsumo: An intelligent transportation system for urban mobility*. LNCS Springer, no. 3473, pp 224-235, 2006.

[4] S. A. Boxill, L. Yu, *An evaluation of traffic simulation models for supporting ITS development*. Technical, Transportation Training and Research, Texas Southern University, USA, 2000.

[5] F. Zhang, J. Li, Q. Zhao *Single-lane traffic simulation with multi-agent system*. IEEE Conference on Intelligent Transportation Systems, pp. 56-60, 2005.

[6] B. Chen, H.H. Cheng *A review of the applications of agent technology in traffic and transportation systems.* Trans. Intell. Transport. Sys., vol. 11(2), pp. 485-497, 2010.

[7] B. Burmeister, A. Haddadi, G. Matylis, *Application of multi-agent systems in traffic and transportation* Software Engineering. IEE Proceedings, vol.144, no.1, pp.51-60, Feb 1997.

[8] J. Gonçalves, *Applying Serious Games to Assess Driver Information System Ergonomics*, Msc. Thesis, Departement of Informatics, Faculty of Engineering, University of Porto, 2012.

[9] J. R. Parker et al., *The Booze Cruise: Impaired Driving in Virtual Spaces* IEEE Computer Graphics and Applications, vol. 29, no. 2, pp. 6-10, Mar. 2009.

[10] J. Gonçalves, C. Olaverri-Monreal, R. Rossetti,*IC-DEEP: A serious games based application to assess the ergonomics of In-Vehicle Information Systems*, Proceedings of the 15th Intelligent Transportation Systems Conference, Anchorage, AK, USA, 16-19 Sep. 2012.

[11] L. Matias, J. Gama, J. Mendes-Moreira, and J. Sousa, *Validation of both number and coverage of bus Schedules using AVL data*, 13th International IEEE Annual Conference on Intelligent Transportation Systems, Madeira, Portugal, 2010.

[12] J. Engstrm et al., *INFORMATION SOCIETY TECHNOLOGIES (IST) Driving performance assessment*, Contract, no. March 2004, pp. 1-149, 2005.

[13] E. Hollnagel and I. Lau *A systemic model for driver-in-control* in Proceedings of the Second International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, pp. 86-91, 2003.

[14] B. Chen and H.H. Cheng *A review of the applications of agent technology in traffic and transportation systems* in IEEE Transactions on Intelligent Transportation Systems, vol. B, issue 2, pp 485-497, 2010.

[15] A. Bazzan, J. Wahle and F. Klugl *Agents in Traffic Modelling - From Reactive to Social Behaviour* in 23rd Annual German Conference on Artificial Intelligence Bonn, KI-99: Advances in Artificial Intelligence, pp 303-306, 1999.

[16] R. Rossetti, R. Bordini, A. Bazzan, S. Bampi, R. Liu, D. Vliet *Using BDI agents to improve driver modelling in a commuter scenario* in Transportation Research Part C: Emerging Technologies, vol. 10, issues 56, pp 399417, 2002.

[17] A. Bazzan *Opportunities for multiagent systems and multiagent reinforcement learning in traffic control* in Autonomous Agents and Multi-Agent Systems, vol. 18, issue 3, pp 342-375, 2009.

[18] R. Sukthankar, S. Baluja, and J. Hancock *Multiple adaptive agents for tactical driving* in Appl. Intell., vol. 9, issue 1, pp. 723, 1998.

[19] P. Hidas *Modelling lane changing and merging in microscopic trafc simulation* in Transporation Research Part C: Emerging Technologies, vol. 10, issue 5-6, pp. 351371, 2002.

[20] P. Hidas *Modelling vehicle interactions in microscopic simulation of merging and weaving* in in Transporation Research Part C: Emerging Technologies, vol. 13, issue 1, pp. 3762, 2005.

SESSION 3

WIRELESS COMMUNICATIONS AND COMPUTER NETWORKS

Protocol for Channel and Gateway Assignment in Single-radio Stub Wireless Mesh Networks

Filipe Teixeira, Tania Calcada, Rui Campos, Manuel Ricardo
INESC TEC, Faculdade de Engenharia, Universidade do Porto
Porto, Portugal
Email: {fbt, tcalcada, rcampos, mricardo}@inescporto.pt

***Abstract*—IEEE 802.11-based Stub Wireless Mesh Networks (WMNs) are a cost-effective solution for extending the coverage of a wired infrastructure. They are formed by single-radio mesh nodes with a single gateway and a single frequency channel which precludes WMN scalability. Using multiple channels and gateways is a solution to overcome the problem, but requires a channel and gateway assignment protocol for automatic and dynamic configuration of the nodes. The Topology Discover and Channel Change (TDCC) protocol has been proposed for this purpose, but lacks support for multiple gateways per channel and only works together with a centralized channel assignment algorithm.**

We propose a Multi-Gateway amendment to the TDCC protocol, named MG-TDCC. MG-TDCC enables centralized and distributed channel and gateway assignment, assuming multiple gateways running in the same or different frequency channels. The centralized approach is compatible with a centralized channel assignment algorithm and manual assignment. The distributed approach considers a built-in distributed algorithm that enables independent channel and gateway configuration. MG-TDCC advantages and foreseen performance improvements are discussed.

***Keywords*—Wireless Mesh Networks, Channel Assignment, Stub Networks, Load Balancing**

I. Introduction

Ubiquitous connectivity to the Internet has become a standard nowadays. To achieve that, IEEE 802.11 Stub Wireless Mesh Networks (WMN) have been proposed for low cost and large coverage networks that are able to extend the wired infrastructure. These networks consider a set of static Mesh Access Points (MAPs) that are interconnected by wireless links in a mesh topology. The static mesh topology means that the topology is only changed by an event, for example, when a node is not responding or when a Centralized Channel Assignment Algorithm (CCAA) or the Network Manager in a Management Tool decides to change the nodes' channel. This has great advantages in terms of control, which is interesting for telecommunication operators. Wi-Fi network Infrastructure eXtension (WiFIX) is a solution for Stub WMN that is simpler and more efficient when compared with other solutions like IEEE 802.11s, using a single-message signaling protocol [1].

Increased number of MAPs will lead to performance decrease due to higher level of interference and possible increasing number of hidden nodes. Using multiple channels can help increasing the network performance, as shown in [2] and [3]. One of the possible solutions is to use multiple radios, one for each channel. However, the usage of more than two radios in the same node, besides the increased cost, leads to severe interference, even if using non-interfering channels [4], [5]. Therefore, in order to have one radio to serve clients, the mesh network is limited to one radio. Keeping a single-radio solution and using multiple channels to increase the capacity of the mesh network demands a channel assignment protocol. Topology Discover and Channel Change (TDCC) protocol [6] is able to discover the links available between nodes, even if they operate in different channels. The topology information is then sent hop by hop and collected at the Master MAP. With this information, the CCAA or the Network Manager can decide the best channel to each node. The decisions are applied to the selected MAPs instantly by TDCC protocol.

In cases where a large number of MAPs are required, using multiple channels may not be enough to decrease the number of MAPs around each gateway and solve the performance problem. This problem is even worse if we consider IEEE 802.11g [7], with only 3 non-interfering channels. Thus we have to consider more than one gateway per channel, in different geographical locations, which demands a combined gateway and channel assignment protocol within the mesh network.

This paper aims to provide the TDCC protocol the ability to support multiple gateways per channel and provide both centralized or distributed gateway assignment of multi-channel Single-radio Stub WMNs. While the centralized approach requires the existence of a Management Tool where topology decisions are performed by the CCAA or the Network Manager, the distributed approach considers an algorithm, running in every MAP, that will enable independent channel and gateway configuration according to a set of metrics. The Multi-Gateway (MG) TDCC proposed on this paper is able to (1) discover all links available between MAPs, independently of the operating channel, (2) collect the topology of the mesh network and deliver it to Management Tool, (3) support multiple gateways per channel, (4) perform automatic channel and/or gateway changes according to hop distance, (5) perform load-balance between different mesh sub-networks, and (6) provide support for manual control over the channel of each node according to the CCAA or the Network Manager decisions. This mechanism helps increasing the overall Stub WMN capacity.

The paper is organized as follows. Section 2 describes WiFIX Mesh Network Architecture. In Section 3, the TDCC protocol is presented. The architecture of the Multi-Gateway TDCC protocol is proposed in Section 4. Section 5 presents a Discussion on the proposed solution. Section 6 concludes the paper and presents future work.

II. WiFIX Mesh Network Architecture

The Multi-Gateway Topology Discover and Channel Change (MG-TDCC) protocol proposed in this paper extends the Wi-Fi Network Infrastructure eXtension (WiFIX) architecture [1]. Based on IEEE 802.1D bridges [8] and a single-message protocol, WiFIX is a simple and efficient solution for Stub Wireless Mesh Networks. This kind of networks aim to extend the current wired infrastructure with a set of static Mesh Access Points (MAPs) that perform multi-hop bidirectional forwarding between the clients and the infrastructure, as shown in Figure 1. Each MAP is equipped with two radios, one devoted to the mesh network formed by all MAPs, and the other one to serve clients.

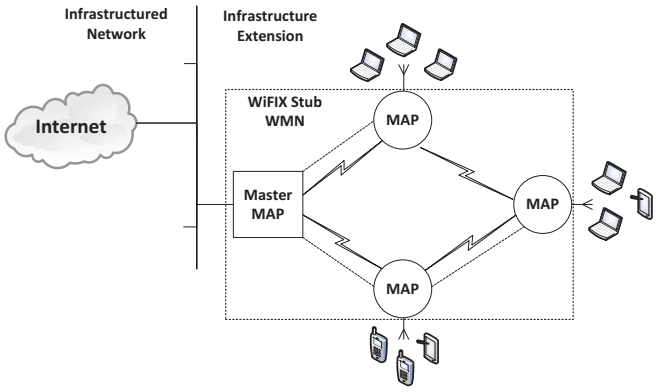


Fig. 1. WiFIX Reference Scenario. Each MAP is equipped with two NICs, one for the mesh network and the other for clients.

The network self organization is achieved by the Active Topology Creation and Maintenance (ATCM) mechanism. ATCM automatically creates a single tree rooted at the Master MAP, which is connected to the wired infrastructure, as seen in Figure 1. This mechanism is supported by a single-message protocol, where a Topology Refresh (TR) message is periodically sent by the Master MAP and forwarded by every Slave MAP. TR messages are sent in broadcast mode and carry the parent address and the distance from the current node to the root (Master MAP). The structure of the TR messages are presented in Figure 2. Every Slave MAP periodically chose the best parent MAP, which is the parent with the least hops to the Master MAP, according to the Distance field of the TR message. This metric, although being simple is effective, shows good results when compared with the radio aware routing metrics presented in [9], which are known to have problems with network instability [10]. The MAC address of the chosen parent is passed in the next TR message in Parent Address field. Therefore, TR messages play three important roles: 1) inform the parent MAP about a new child; 2) announce the MAP presence to other MAPs and 3) announce the Master MAP. The encapsulation of the Ethernet frame inside the 802.11 frame is performed by Ethernet-over-802.11 (Eo11) tunneling mechanism. Eo11 allows to store the original source and destination addresses, freeing two MAC addresses for forwarding the frames from MAP to MAP using the simple learning mechanism of the 802.1D bridges.

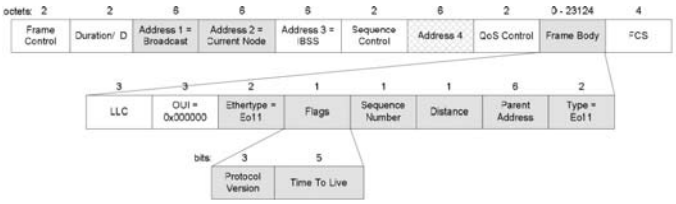


Fig. 2. Topology Refresh (TR) Message structure. TR messages are sent by the master MAP and forwarded by every Slave MAP.

III. TDCC Protocol

Topology Discover and Channel Change protocol was designed to allow multi-channel operation in WiFIX Stub Networks. To keep only one radio interface in the mesh network, a channel assignment protocol is needed. In [11] there are shown 4 different types of protocols and architectures for single-radio channel assignment in WMNs. The first one, Dedicated Control Channel protocols, require that one channel must be reserved for control packets, wasting 1/3 of the available orthogonal channels in 2.4 GHz IEEE 802.11b/g/n networks. Hopping protocols and Split-Phase protocols require complex synchronization mechanisms that are not compatible with the current IEEE 802.11 standard [7], [11] . In the first case, mesh nodes hop between channels according to a pattern. In the second case, the time is divided in cycles with a control phase and a data phase. Receiver-fixed protocols consider the existence of a fixed quiescent channel associated to every node. When the node wants to send data to another node, it changes to the quiescent channel of the destination node, returning to its quiescent channel as soon as the transmission ends. After that the node is free to receive data from other nodes. Despite the increased bandwidth due to broadcast transmissions, Receiver-fixes protocols are compatible with IEEE 802.11 networks and are easy to implement.

TDCC is a Receiver-fixed protocol that is responsible to collect in the Master MAP the links available between all MAPs, independently of their operating channel, without introducing additional messages to the WiFIX solution. The free space of TR messages from original WiFIX ATCM mechanism were used to store topology data. Each TR message received is retransmitted in each channel, which allows that nodes tuned in other channels receive the TR message and know about neighbors tunned in other channels. With the information about the topology of the mesh network, the CCAA or the Network Manager can fine-tune the operating channels of the MAPs in the Management Tool and assign new channels to each MAP. The protocol will process those decisions and apply the channel changes towards the tree. In Figure 3, the reference scenario of TDCC protocol is presented, where two different trees can be seen, rooted at Master MAPs in different channels.

TDCC operates in two modes: Topology Discovery mode and Channel Change mode. In the Topology Discovery mode, the protocol collects information about the links available between MAPs and report that information to the Master MAP. In every TR message received, the MAC address and quiescent channel is stored in a neighbor list. Then, each node will append its neighbor list to the TR received by its child nodes and send a new TR message with that information. Through this mechanism, the Master MAP will have the knowledge

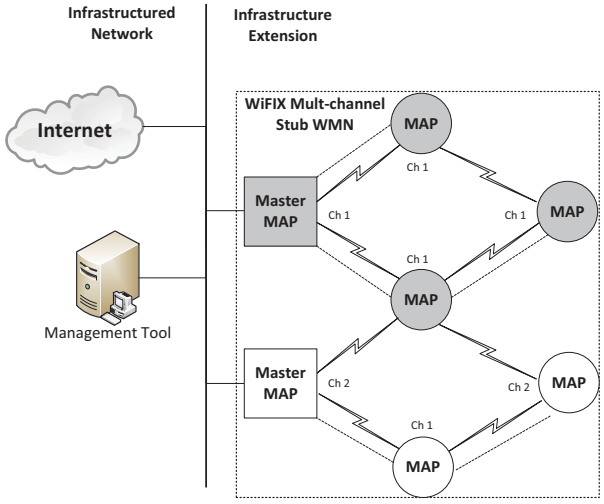


Fig. 3. Reference scenario of TDCC protocol, with two trees in two different channels. The Management Tool receives the topology of the mesh network and enable manual channel changes performed by the CCAA or by the Network Manager.

of the network topology after $h + 1$ TR messages, being h the number of hops. Signaling packet loss is avoided with an hysteresis period. During that period, T_{Upd}, each node will collect the topology information from the child nodes and only after that period the information is sent to the parent node. This avoids premature topology changes that would occur if TR messages are lost in case of collisions or momentary interference. With this mechanism set, the convergence time is defined by $T_{Upd} * (h+1)$. The Channel Change mode applies the decisions on the topology set by the CCAA or by the Network Manager. The TR message sent by the Master MAPs contains the MAC address of the nodes that need to change and their new channels. As every node retransmits this message from the ATCM mechanism, these orders will be received by every node in the mesh network and the nodes required to change perform this operation instantly.

IV. MG-TDCC Protocol

Multi-Gateway Topology Discovery and Channel Change (MG-TDCC) protocol aims to provide TDCC protocol the support for multiple gateways in the same channel. This will avoid situation where the high number of nodes around one gateway, would lead to starvation of the MAPs which are far from the Master MAP, making the communication between them impossible or at extremely low rates. With MG-TDCC there is a division of the stub WMN, not only in terms of channels but also in several gateways per channel. This case is specially interesting if the gateways are geographically separated, and not concentrated in one specific location. In scenarios like football stadiums, airports or shopping centers, MG-TDCC together with the WiFIX solution provides a flexible, scalable and redundant network with low deployment cost.

A. Master MAP identification

Having more than two gateways in the same channel requires that the different messages generated by the multiple Master MAPs should be distinguished. Otherwise a slave MAP can not identify which Master MAP was the TR related to. A new format for the Topology Refresh (TR) message was defined to overcome this limitation: besides the Sequence Number, the MAC address of the Master MAP should be sent in the TR message. The dark gray blocks of Figure 4 show the differences introduced by MG-TDCC, while the light gray show the original TDCC message. Every Slave MAP will broadcast every TR message received, increasing the distance to the Master MAP.

B. Automatic Channel Change

Changes in Channel Change mode were performed to optimize the solution for the multi-gateway scenario. In TDCC, changes in the MAP channel were performed by the CCAA or by the Network Manager. The only exception was when the node become isolated in one channel, due to the failure of its parent or by an erroneous channel assignment. In MG-TDCC protocol, the node will change channel if there is a parent in another channel with less hops than the current parent. This automatic channel change can be disabled if the CCAA or the Network Manager assigns a fixed channel, order that is sent to the Master MAP and propagated by the ATCM mechanism in the TR message payload. To re-enable this feature, CCAA or the Network Manager should assign the channel 0 to that MAP. The automatic channel change feature will lead to the creation of a topology with the minimum possible hop distance.

C. Traffic-aware decisions

In cases where there is a break-even between different possible parent MAPs with the same number of hops, the decision on which channel and/or gateway to choose can be based on the tree load. MAPs should choose the parent with the least Tree Load. To calculate the traffic load of each tree, every MAP will measure the traffic sent to the clients connected to it during a certain period of time. This value is reported to the parent node through the Load field of the TR message (Figure 4). The Load field of every child node is collected and appended to the Topology Data of each child, which is passed to the parent node in the next TR message. This mechanisms allow the Master MAP to have the information about the load generated by every MAP. The tree load is calculated by the Master MAP, weighting the traffic according to the hop distance and the Tree Load field of the TR message sent by the Master MAP is updated. If the node is already associated to one parent and there is another parent with the same number of hops to the Master MAP, each Slave MAP should evaluate if changing to another parent with the same number of hops to the gateway implies a Tree Load greater than the current Tree Load. To do it, each MAP will sum the Tree Load to the traffic generated by the own MAP multiplied by the hop count, as every hop implies resending the same packet After that, it compares the new Tree Load with the current Tree Load. The MAP should change only if the new Tree Load is lower. If the MAP changed from one tree to another without this verification, it could cause network instability due to constant parent changes. This approach is similar to the load-balance protocol presented in [12].

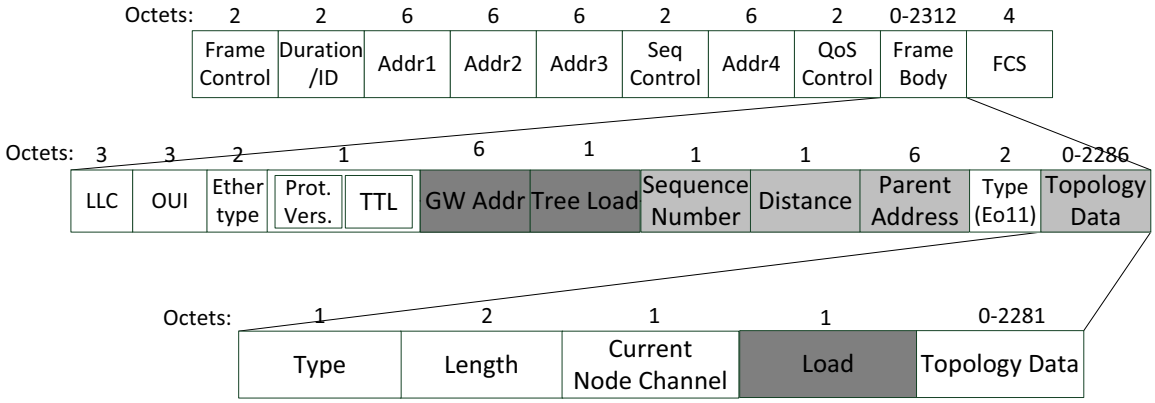


Fig. 4. New TR message structure to support multiple gateways and the tree load ratio.

D. Validation

In order to validate the MG-TDCC protocol, we propose an implementation of this protocol embedded in WiFIX daemon. WiFIX daemon runs in any Linux system and is placed between the Linux Bridge and the 802.11 Network Interface Card (NIC). The implementation should be tested in a 12 node testbed as shown in Figure 5, where the circles represent the desired node location and the network topology, while the pins represent the actual geographical location. This testbed allows multiple topologies of the mesh network, including 4 Master MAPs: 2 MAPs in 2 different channels. This allows to test the performance of the network when automatic channel assignment and gateway choice is selected, but also to evaluate the performance on certain mesh topologies.

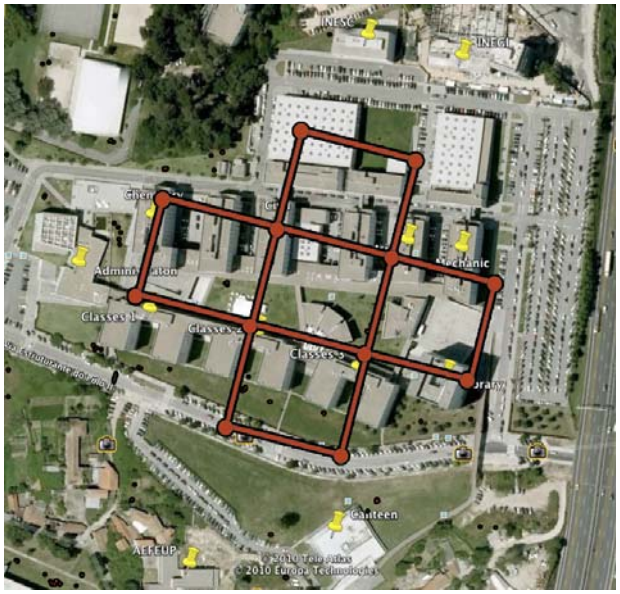


Fig. 5. Proposed 12-node Testbed for MG-TDCC validation. The circles represent the desired node location and the network topology, while the pins represent the actual geographical location.

V. Discussion

The improvements introduced by MG-TDCC allow the load-balance of the MAPs among different gateways and different channels, with promising results.

A. Convergence time

The Topology Discover mode of TDCC, which is responsible for collect the topology of the WMN is not affected, even if we consider having multiple gateways. Each node creates a table of neighbors and their channels and send it to the parent node. During a certain period of time (T_{Upd}), the parent node collects the topology information from all child nodes and will append that information to the next TR message sent. Even considering multiple trees in the same channel, this mechanism remains the same. Parent nodes will only collect the topology information from the messages with the Parent Address equal to the node's MAC. Therefore, according to [6], the time required to report changes to the Master MAP is small enough for the static Stub WMNs considered in this paper. Figure 6 represents the parent MAP selection when multiple gateways are present. If we consider the scenario of the first three nodes of Figure 6 and the case where the MAP 2 joins the network, Master A will take 20 seconds to be aware of the association between MAP 1 and MAP 2. This results consider the time interval between TR messages of 2 seconds, and the hysteresis mechanism allowing up to 50% signaling loss. In case MAP 2 is turned off, Master A will receive that information after 15 seconds. The convergence time observed is adequate for the static Stub WMN scenario, keeping the overhead introduced very low while allowing tolerance for signaling loss.

B. Channel change

MG-TDCC does not introduce any changes to the Channel Change mode of TDCC protocol. This means that MG-TDCC should perform almost instantaneous channel changes. The time between the order from the CCAA or the Network Manager and the channel is changed depends mostly on the network delay [6].

C. Multiple gateways in the same channel

Figure 6 shows the case when a new node, node 3, appears in the mesh network. Node 3 will receive two TR

messages, one from MAP 2, connected to Master A and other from MAP 4, connected to Master B. With the two messages, Node 3 will select the MAP with the least hops to the root of the tree, which is MAP 4 in this case. Choosing MAP 4 as parent, MAP 3 will rebroadcast the TR received from MAP 4 and update the Parent Address with the MAC address of its new parent and increase the distance by 1. MAP 3 will also rebroadcast the TR message from MAP 2, increasing the distance, but keeping MAP 4 MAC in Parent Address field. This shows the capability of MG-TDCC to cope with multiple gateways per channel.

Fig. 6. Parent MAP selection when multiple gateways are present. When MAP 3 is turned on it will choose MAP 4 as parent as it is the MAP with least hops towards one of the Master MAPs.

D. Automatic channel changes

If we consider that the light-gray MAPs in Figure 6 are tunned in Channel 1 and dark-gray MAPs are tunned in Channel 6 of IEEE 802.11b/g networks, MAP 3 when turned on will receive both TR messages from MAP 2 and MAP 4 and will tune itself to Channel 6 as MAP 4 is only 1 hop from the Master B. This automatic reconfiguration of the nodes should improve the network throughput and decrease delays, and is an important feature when failures or interferences forces changes in the Stub WMN topology.

E. Traffic-aware gateway changes

In Figure 7 is presented an example of the traffic aware decisions, where the MG-TDCC will perform load balancing according to the Tree Load. Here, MAP 2 is 2 hops away from both Master A and Master B. Its Load is 5 and the Tree Load of tree formed by the Master A is 30 and from Master B is 10. If the parent of MAP 2 would change to MAP 3, the new Tree Load would be $10+2*5=20$. As the new Tree Load of the tree associated to MAP 3 is smaller than the Tree Load of the MAP 1, MAP 2 would automatically change if automatic change channel is enabled, which is expected to improve the network performance.

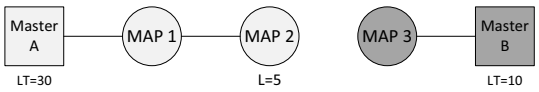


Fig. 7. Load balancing depending on the Tree Load. MAP 2 will change its parent from MAP 1 to MAP 3 if the the traffic Master B is lower than Master A, even considering the increase of traffic induced by MAP 2.

VI. Conclusion

In this paper we propose an amendment to the Topology Discover and Channel Change (TDCC) protocol to support multiple gateways per channel in WiFIX Single-radio Stub Networks. Together with WiFIX, MG-TDCC protocol is able not only to discover the topology of the mesh network, independently of the channel each node is operating, apply channel changes decided by a Centralized Channel Assignment Algorithm or the Network Manager in a Management Tool, but also to automatically balance mesh nodes between multiple channels and multiple gateways per channel. MG-TDCC is interesting for deploying a large WMN for covering a large area and a large number Mesh Access Points as it increases the scalability of the traditional mesh solutions for Stub Networks, combining both centralized and distributed mechanisms for channel and gateway assignment. With this protocol, besides the manual decisions that can be applied to this network, the network has the capability to adapt itself according to the current topology and the traffic that each node is carrying in a certain period of time, providing a complete and effective solution for infrastructure extension with both centralized and distributed topology control.

The future work out coming from this work includes the full implementation of the protocol and a comprehensive set of tests in a 12-node testbed installed at FEUP buildings, which is expected to show the benefits MG-TDCC protocol.

References

[1] R. Campos, R. Duarte, F. Sousa, M. Ricardo, and J. Ruela, "Network infrastructure extension using 802.1D-based wireless mesh networks," *Wireless Communications and Mobile Computing*, vol. 11, pp. 67–89, Jan 2011.

[2] M. K. Marina, S. R. Das, and A. P. Subramanian, "A topology control approach for utilizing multiple channels in multi-radio wireless mesh networks," *Computer Networks*, vol. 54, no. 2, pp. 241–256, Feb. 2010.

[3] P. Kyasanur and N. H. Vaidya, "Capacity of multi-channel wireless networks: impact of number of channels and interfaces," in *Inter. Conf. on Mobile computing and networking (MobiCom'05)*, Germany, 2005.

[4] A. Dhananjay, H. Zhang, J. Li, and L. Subramanian, "Practical, distributed channel assignment and routing in dual-radio mesh networks," *SIGCOMM*, vol. 17-21, pp. 99–110, August 2009.

[5] J. Robinson, K. Papagiannaki, C. Diot, X. Guo, and L. Krishnamurthy, "Experimenting with a multi-radio mesh networking testbed," *Proceedings of the First Workshop on Wireless Network Measurements (WiNMee 2005)*, April 2005.

[6] F. Teixeira, T. Calçada, and M. Ricardo, "Protocol for centralized channel assignment in wifix single-radio mesh networks," in *3rd International ICTS Conference on Mobile Networks and Management (MONAMI 2011)*, 2011.

[7] IEEE 802.11 Work Group Part 11, "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications," Tech. Rep., Jun 2007.

[8] IEEE 802.1D Work Group Part 11, "IEEE standard for local and metropolitan area networks: Media access control (MAC) bridges," Tech. Rep., Jun 2004.

[9] IEEE 802.11s-2011/ amendment to standard IEEE 802.11, "Mesh networking," Tech. Rep., Jun 2011.

[10] R. G. Garroppo, S. Giordano, and L. Tavanti, "Implementation frameworks for IEEE 802.11s systems," *Computer Communications*, vol. 33, pp. 336–349, Feb 2010.

[11] J. Crichigno, M.-Y. Wu, and W. Shu, "Protocols and architectures for channel assignment in wireless mesh networks," *Ad Hoc Networks*, vol. 6, pp. 1051–1077, Oct 2007.

[12] J. So and N. H. Vaidya, "Load-balancing routing in multichannel hybrid wireless networks with single network interface," *IEEE Transactions on Vehicular Technology*, vol. 55, pp. 806–812, May 2006.

Overview of Integrated Network for Oil Pipeline Monitoring

O. Aboderin
Faculty of Engineering, University of Porto, Porto Portugal
mpt12005@fe.up.pt

Abstract – **The search for a near perfect solution to monitor oil pipelines leakages and breaks, have led to many research on different technologies that can be used to achieve this. In actual fact, there have been many cases of both non-intentional (naturally occurring) faults and deliberate or intentional (man-made) attacks on the pipelines, which often results in enormous loss in terms of the products and the ensuing inferno. It is also evident that using many of the externally based Leak Detection Systems technologies independently is not achieving the desired aim. Presently, independent networks of Wireless Sensor Network, Fiber Optics, Long Range Camera, Satellite, and Unmanned Aerial Vehicle have been used to monitor the pipelines, but events in the past years have revealed that they are not sufficient singularly for the task. It is to this end that this paper looks into the need for integrated networks and gives the overview of integrating at least three of these technologies for thorough surveillance of the pipelines. The proposed configuration will be linked with mobile terminals and control stations for effective and timely responses, when utmost needed and each technology in the system will work to complement the other. Costs and energy required were also taking into consideration in the proposed combined surveillance technology design.**

Keywords: pipelines, Leak detection Systems, integrated networks, control stations, mobile terminals.

I. INTRODUCTION

There is no doubt about the indispensable use of pipelines for effective transportation and distribution of water, oil, gas, slurry and chemicals across many nations of the world [1], [2], [3]. Thus, defects in these systems will not only lead to huge financial loss but will also results in environmental pollution which can cause epidemic and fire outbreak that can lead to loss of life and properties [3]-[5]. These pipes are faced with non-intentional threats like ruptures, breaks, damages and leakages, which can be due to; aging of the systems, human errors in operation and maintenance, natural disasters like earthquakes and volcanoes. But oil and gas pipelines face more threats from deliberate sabotage by vandals for illegal tapping or for terrorism [3]-[6]. Leak Detection Systems (LDS) has been used over the years to assist pipeline controllers to detect and localize leaks, report instantaneously, provide alarm when necessary and display other related data to the control station. It is divided into externally and internally based. The internally based LDS make use of field instrumentations like pressure, temperature or flow rate to monitor pipelines internal parameters. On the other hand, externally based LDS monitor the external parameters of the pipelines by using the following field instrumentations [7], [8], [10]:

- Sensor Networks;
- Long Range Cameras or Infrared Radiometer;
- Fiber Optics Cables;
- Unmanned Aerial Vehicle;
- Satellite Networks.

The concern of this paper is to look into the externally based LDS. In the recent times, many works have been done about using these technologies independently for monitoring and surveillance of pipelines, but the output have shown that independent technology is not enough for proper monitoring and surveillance in many of the oil riched states. This paper is therefore giving an overview of definite solution to problems of pipelines breaks, by full monitoring to ensure smooth delivery of the products from the oil well/rigs (onshore or offshore) to refineries or manufacturing plants and from these plants to the marketers and homes for final distribution. The proposed network will combine three technologies earlier mentioned, that have been investigated and certified that it can provide monitoring of the pipelines to form an integrated multi-system monitoring network. In the build-up, cost of providing this will be considered and the energy required to run the technology. It is worthy of note to state that monitoring and surveillance are used interchangeably throughout this paper. Also, the rest of the paper is organized as follows. Cases of vandalized pipelines as well as leakages are considered in section II. Section III considers why it is almost impossible to use the technologies independently in providing proper surveillance and monitoring. The propose architecture of the integrated technology is discussed in section IV. Section V looks into the cost both from continuing pipeline breaks and providing the integrated technology. The paper concludes in section VI.

II. CASES OF PIPELINES BREAKS AND LEAKAGES

Many countries in the world have different but related stories to share about the pipeline leakages in their country. For instance, in Nigeria it was reported that within the last decade, about 16,083 pipeline breaks were recorded, of which only 398 occurred as a result of aging and rupture of the pipelines, the rest 15,685 came as a result of unpatriotic vandals for illegal siphoning. In all, about $1.1billion was incurred on product losses and pipeline repairs [4], [5], [9], [11]. Apart from the huge financial loss, there were also loss of life in many of the incidents through the inferno resulted from scooping of the highly inflammable liquid. The aquatic lives were not spared as oil spillages released into rivers led to their death. This is not limited to this African country as similar occurrences were recorded in Columbia, Mexico,

USA, Belgium, Kenya, China and the Middle East. In Colombia there were over 900 attacks on the Cano Limon oil pipeline, in 2002 and this led to losses of around 2.5 million barrels of crude oil [4], [9], [11]. Earlier in 2001, the same pipeline was out of service for 266 days in the year due to the blown up of pipeline, which is up to 170 times within the year. The situation in Mexico is combination of illegal tapping and terrorism. In July 2007, Leftist Popular Revolutionary claimed responsibility for bombing of a gas pipeline and thus stopped the flow of natural gas from Petroleos Mexicanos (PEMEX) pipelines in Mexico and in December 2010 there was an explosion on their (PEMEX) oil pipeline at the pumping station in San Martín Texmelucan de Labastida in central Mexico, in this incident 2 people were killed and more than 50 injured. The explosion is believed to have occurred as a result of illegal tapping by oil theft [4], [9], [11]. The case in Kenya is quite different as systems failure led to leakages along the pipeline of Kenya Pipeline Company (KPC) in 2011 and this led to deadly explosion in which about 100 people were killed in the ensuing inferno [5]. In July 2010 in the northeastern port city of Dalian, China, two pipelines exploded releasing thousands of gallons of oil into the nearby harbor and Yellow Sea as shown in fig. 1 [9], [12]. These are just few examples from numerous pipelines breaks across the world. In all these, who knows if there are early detection of the ruptured pipelines, the unprecedented loss, the deaths, the inferno, and the spillage would have been prevented.

III. Defects In Using Each Technology Independently

For effective and efficient monitoring, it is not really possible to use this technology independently to monitor oil pipelines. Each of them is thus considered with their strength and weaknesses.

A. Wireless Sensor Network

The ease of deployment of WSN that does not include laying of cables, the low cost involved and the secured communication networks are advantages that makes WSN useful in monitoring pipeline [13]. A typical sensor node consist of four basic components; a sensing unit, a processing unit, a communication unit and a power unit. The sensing unit comprises of sensors and Analog to Digital Converter (ADC). Here, the sensor observe the phenomenon and generate analog signals based on its observation, ADC however, converts the analog signals into digital before feeding this to the processing unit. The processing unit has a microcontroller, which supply intelligent control to the node. Transmission and reception of messages is performed by the short range radio in the communication unit. The power unit gives life to the node providing the needed power by the sensor node and this is supplied by the battery. Good architecture and configuration of the network can also provide real time monitoring, by collecting data and giving updates of events on and around the pipeline, but if not properly configured it can also give false alert [14]. Though, with Global Positioning Systems (GPS), WSN can be used to determine the actual location of the disturbances to the pipelines but this is not enough to get field images, there will be need for other technologies that can provide this service. Continuous availability of the network is another important factor in the use of WSN as failure of sensor nodes may sometimes result in failure of the entire network.

B. Optical Fiber

Optical fiber is known to have high accuracy in communication signal transmission, this responsible for it applications in many communication systems. A single mode distributed fiber can be configured to monitor pipeline in real-time through Fiber Optic Sensor (FOS) techniques [2], [15].

Fig. 1. Oil Spillage in China's pipeline explosion [12]

It is good to note that role of GPS in WSN is equivalent to what Optical Time Domain Reflectometry (OTDR) does in optical fiber, when there is need to know precise location. Monitoring along the pipelines is possible for more than 30km length using this technology to detect leakages, ground movement, structural health monitoring and third party activities [15]-[17]. Despite these advantages, it is still a known fact that the technology is not all alone enough to do the job, just like the pipeline, fiber itself can be vandalized, especially when it is a deliberate or intentional attack on the pipeline. Recently a communication giant in Sudan expresses their concern over the manner at which this technology suffered a great deal from the hands of the vandals. Similar incidents were reported by AT&T in US and the cost of repair or replacement is very high as well, this will add to the difficulty in carrying out the repair especially for the pipelines in difficult terrain [18].

C. Unmanned Aerial Vehicle

UAV systems fully equipped with; remote mapping, aerial photographing, On Board Infrared Equipment (OBIE) and communication tools and effectively networked to control station(s) will deliver good monitoring and surveillance in real-time and low cost compared to the satellite [19]. This will also reduce the latency (round trip time delay) that would have been experienced with satellites. Also by building an interfacing ends between the WSN and UAV, the later will

assist in delivering the acquired information by WSN to the control station. The vehicle delivers a near earth aerial mapping and surveillance majorly in the areas where deployment of any other technology to monitor the pipeline might be extremely difficult [20] [21]. Besides using it for pipeline, it has other applications, depending on the onboard payload of such system. Fig. 2 shows the various heights of UAV systems with the corresponding coverage lengths and the implication on the network.

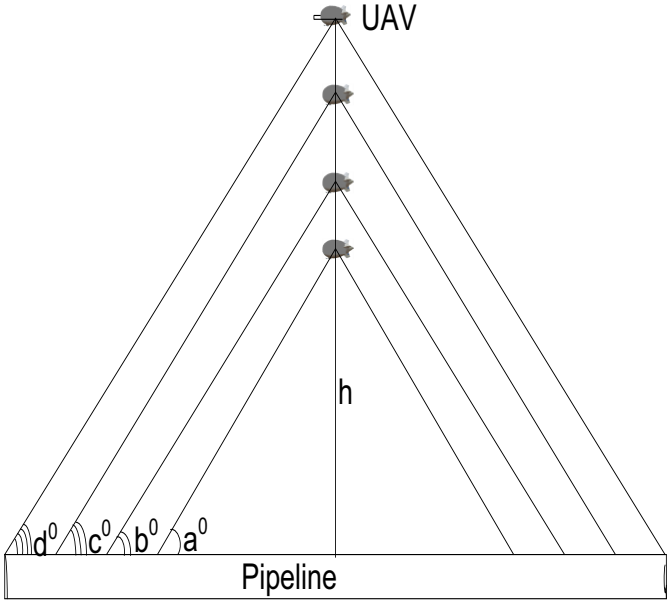


Fig. 2. UAV heights and coverage

Mathematically from similar triangles it gives that;

$$a = b = c = d \tag{1}$$

Where h is the altitude and a, b, c and d are the corresponding elevation angles as the altitude of the UAV increases. This implies that with higher altitude, the system will be able to have more coverage areas but diminish images. It is the same analogy that is employed in distinguishing satellites launched into various orbits. From this analysis, it is also obvious that UAV alone cannot be used single handedly because of the limitations in the height, coverage areas and the acquired images. Also the number of hour it can endure when in the air is another disadvantage that makes full dependent on this technology not worthwhile for monitoring and surveillance.

D. Satellite

Satellite is a highly dependable network that can be used for pipeline surveillance. It started with acquisition of information about an object of considerable distance without any physical contact otherwise known as remote sensing [20], [22], [23]. Then, the reliance was on low resolution black/white aerial photograph. But the advancement in modern technology has seen the spectroscopy techniques evolved through panchromatic to multi-spectral to hyper-spectral and presently ultra-spectra image resolution, with which it is possible to clearly differentiate images on the ground [22]. Likewise the technology also evolved from using Low Earth Orbit (LEO) satellite to Synthetic Aperture Radar (SAR) satellite that is capable of penetrating thick clouds and also very useful for night coverage [20]. But, it will be impossible to get real-time communication through this medium. Also, the extremely high cost of building and launching satellite, equipping ground receiving station, continuously updating the useful software shows that though a good network, it will be too expensive for the task.

E. Long Range Camera

The ability of camera in surveillance technology have been proved in many cities of the world where it has been used for monitoring and surveillance on; major highways, in public places, along the streets and within buildings. Camera can also be used to monitor pipelines for leak detection or human encroachment or third party activities that can responsible for intentional pipeline breaks. Indeed by using IP cameras will ensure that images along the pipeline are delivered instantaneously to the appropriate quarters. This is done by mounting it on a high mast and integrating it with existing networks. But if UAV is equipped with photographing tools, it will be preferable to using this technology, as assailants can intentionally destroy it. With this, depending on this method alone to monitor pipelines may jeopardize the motives of having good surveillance systems along the pipelines. Nonetheless, without attacks, it is a dependable network.

IV. Architecture Of The Integrated System

Considering the inability of the technologies to successfully used for monitoring, thus this integrated network is proposed. The architecture will be varied depending on the requirement of the user of the integrated technology, which might be based on the availability of funds. Fig. 3 shows the coordinated system that will have a WSN, Fiber Optics, and UAV working together to monitor pipelines. Other integrations that can be considered are all displayed in table 1, which in all has WSN, Fiber optics, Satellite, UAV and Camera combined together in the group of three stating the services, advantages, disadvantages and remarks. It is evident that some users will prefer the use of WSN, fiber optics and UAV; others might be interested in another configuration entirely. Thus, camera and satellite were in other configurations to ensure that irrespective of user's requirement, the proposed integrated networks will be good enough to handle it. Though the satellite will work on store and forwarding principle, but will be complementing others that will be programmed for real-time mission. Also, due to the advance imaging techniques earlier mentioned, satellite's inclusion in any combination will be of great advantage in getting useful ground imaging of pipeline leakages and vandals.

For the integrated system in fig.3, the configured sensor nodes of the WSN will employ Category 1 WSNs (C1WSNs) that uses dynamic routing protocol, where each node is configured to acts as an independent router and can have more than one single radio hop from the forwarding node [14], [24]. The source (SN1 & SN2) and sink (SN3) nodes will interoperate with each other through this mesh-based system and each sensor nodes will be equipped with Global

Positioning System (GPS) because location information is required. In this design, sensor nodes will communicate through a unique channel, and this channel has the characteristics of ensuring that at any given time only a single node can transmit message. To this end, Media Access Control (MAC) protocol will be established among the nodes to regulate access to the shared wireless medium. This is implies that communication between the nodes will be through lower sublayer of Data Link Layer (DLL) of the Open System Interconnection (OSI) Reference Model. DLL, also known as layer 2 is subdivided into lower and the upper sublayer and the upper sublayer known as Logical Link Control (LLC). With this, probability of message collision will be close to zero and there will be less retransmission of messages between the nodes, which in turn will save energy that will have been wasted in the retransmission. In this configuration, the source nodes will transmit at regular time interval to SN3, and SN3 will transmit to the Management Station (MS), to the handheld devices or Mobile Terminal (MT) of the security personnel and to the UAV.

Proposing an Unmanned Aerial Vehicle (UAV) in this combined network is to ensure that the cost of providing the solution is not too exorbitant. Actually, this is substitute to using satellite owing to the cost of building and launching of the latter. Also, considering the fact that a near orbit (Low Earth Orbit) satellite is the one that is appropriate for this task, then there will be need for more than one (constellation) in order to provide a real time monitoring which is needed. The UAV is tasked with responsibilities of; communicating with other networks in the system, providing the aerial surveillance coverage and interfacing between the field and the Management Station (MS) or Central Management Station (CMS). UAV will also be relied upon for taking instantaneous photographs, which will have combination of staring, pan, zoom and tilt techniques [25]. Apart from communicating with SN3, linking MS to CMS, UAV will still be able to do aerial mapping, and notify the CMS for an immediate action. In rare cases, it can also be used to aid deployment of sensor nodes, which can be deployed in attended or unattended mode. Here UAV followed a predetermined trajectory (pre-programmed flight pattern) and drops the nodes at predetermined interval at predetermined locations [20]. This will make the deployment faster and easier, but it is necessary to check the nodes, as it is obvious that some of the nodes may become damaged during deployment. It is worthy of note to state that there will be need for proper coordination between UAV and the manned aircraft and this can be done in two ways

1. Declaration of the areas involved as a non-fly zone to the manned aircraft by authority concerned;
2. Proper regulation in conjunction with control towers of airport to ensure that manned aircrafts fly at an altitude far higher or far lower than that of the UAV.

According to Future Fiber Technologies (FFT), distributed fiber optics has an advantage of monitoring up to 50km length pipeline in real-time by employing Raman and Brillouin scattering effect [15], [16]. These are the most developed FOS technologies and they made use of a nonlinear interaction between the light and silica material which is made of fiber. Light at a known wavelength is seen to have a small amount scatter back at every point along the fiber. These scattered lights contain Raman and Brillouin components which are at different wavelength to the original wavelength called Raleigh wavelength. In order to measure, these components and specialized OTDR are used. Short laser pulses are sent to the fiber by measuring devices and the responses is analyzed in time-distance related to the reflected signals with respect to frequency and amplitude of the desired scattering effect, with which measuring strain and temperature is made easy [2], [15]. Also the fiber optics will communicate with MS; this will also be providing a real-time update along the pipelines.

The Management Station will be the sub-control and sub-coordinating unit of the integrated network, with which an interface will be created between the station and each of the technologies in the network. The length of the pipeline determines the number of MS in the network. When any of the MSs receives message from the field parameters, it will send this instantaneously to the MT of the security personnel within their domain and also through UAV or Very Small Aperture Terminal (VSAT) to the CMS. The VSAT link will be secured using Point to point Tunneling Protocol (PPTP) of the Virtual Private Network (VPN) to wade of any form of attacks from hackers. The proposed PPTP is DLL protocol and is preferred to Layer Two Tunneling Protocol (L2TP) and Internet Protocol Security (IPSec) because of its compatibility with many operating systems.

At the CMS, a Supervisory Control Data and Acquisition (SCADA) system will be installed. This is a computer based system that will be used to monitor, process, transmit and display pipeline data in real-time at the CMS. Though the system on its own can be used directly to monitor pipeline, but in this case, it is been proposed to support coordinate the activities of the integrated network. With this, real-time data will be received from the field through Remote Terminal Unit (RTU) and subsequently be sent to appropriate mobile devices for immediate action. Apart from communicating with MS and field parameters, CMS will also create database of MTs of the law enforcement officers and other concern personnel (Pipeline Controllers) and send it accordingly to the MS. As said earlier MS is tasked with responsibility of communicating with MTs by sending message when utmost required. Thus, it is expected that MTs will have multiple communication links that will give precise location of either rupture or breaks. The mobile device will use GPS to update CMS of it particular location, this will assist the later in updating list sent to each MS at regular time interval. CMS will also be required to request for updates of the message sent to MTs by MS, with which it will decide whether to send a reminder to the MTs, or to even send to other MTs that have changed their location. The updates at the MS and CMS might sometimes be sent repeatedly, when there is no fault with any of the technologies in the network. It is thus expected that irrespective of the faults, there will be updates delivered to MS and CMS at the required period.

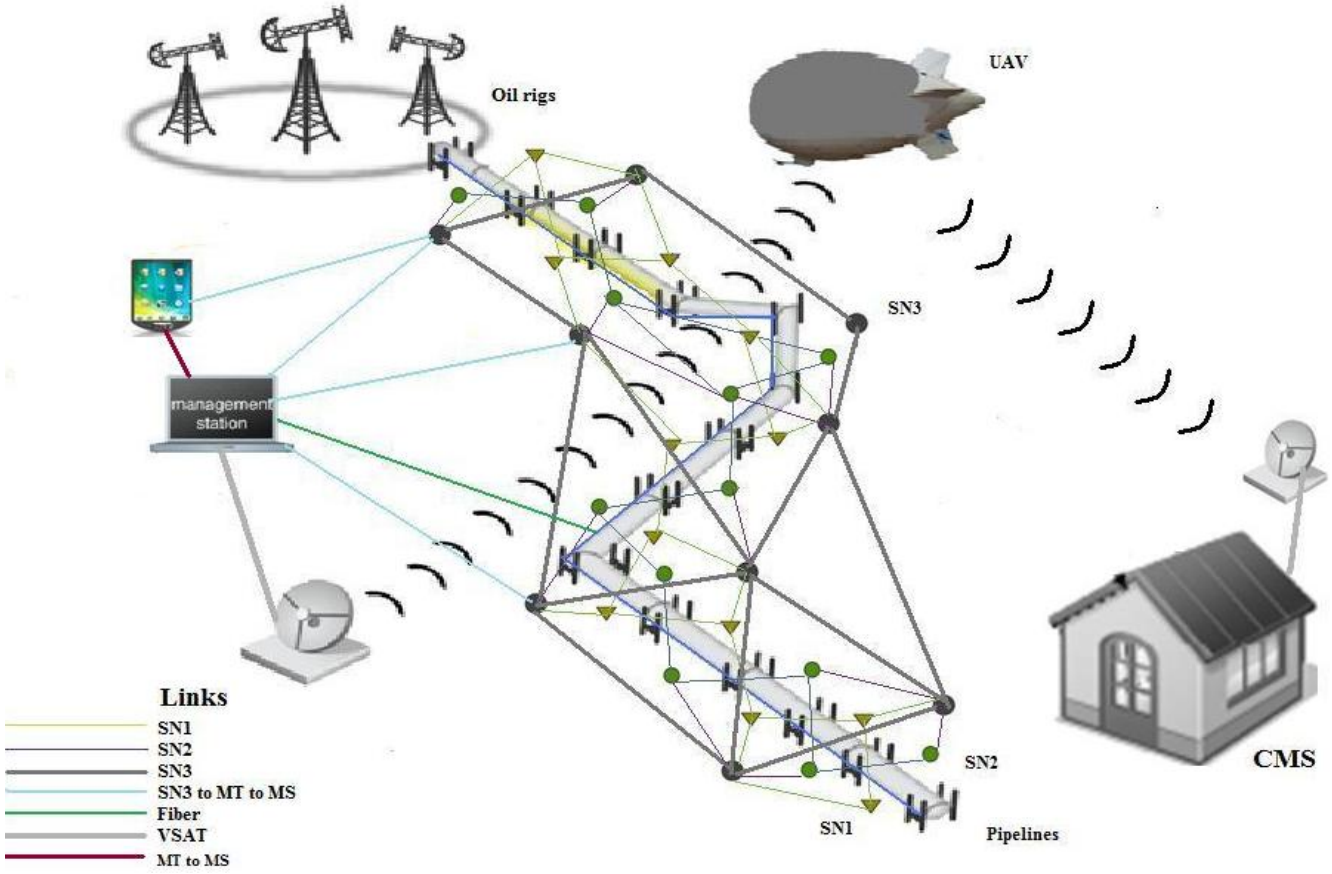


Fig. 3. Architecture of the Integrated Network

TABLE I. Four proposed integrated networks for real-time monitoring of pipeline

S/No.	Integrated Networks	Services	Advantages	Disadvantages	Remarks
1.	Fiber optics + WSN + UAV	Real-time, early intruder detection, aerial photograph.	Low maintenance cost.	Expensive, fiber installation must be right-of-way.	This seems to be the best combination.
2.	Satellite + WSN + Camera	Support store and forwarding, image processing, real-time, early intruder detection.	Relatively low maintenance cost.	Highly expensive, satellite suffers from urban congestion, Camera can be easily attacked.	Cost of adding satellite makes the configuration unaffordable.
3.	WSN + Camera + UAV	Real-time, early detection.	Low maintenance cost.	Camera can be easily attacked.	Camera and UAV can be used in similar manner.
4.	Satellite + WSN + Fiber optics	Store and forwarding, image processing, Real-time, early intruder detection.	Good images, low maintenance cost.	Extremely expensive, satellite suffers from urban congestion, fiber installation must be right-of-way.	Though a good combination, but the implementation cost will be too high.

V. Cost And Energy

Indeed the cost of providing this network of integrated monitoring system will be very high, taking into consideration cost of each network. But when compared to those lost on annual basis to the ruptures and vandals, and also considering that these technologies are more of self-sustaining, then this combination is worth-while. For instance, if country loss $1.1 billion in ten years, it is obvious that 10% of this amount couldn't have been used to provide the needed integrated surveillance technology over the same period of time.

The main network where energy concern is taking into consideration is WSN and it is proposed to draw energy from solar systems. In order to further conserve the energy in the design configuration for the WSN, SN1 and SN2 are designed to be hot redundant to one another. This mean only (SN1) one will be listening at a time, while the other (SN2) sleep and it is designed that they will interchange automatically should there be any fault in the one that is working. It is also expected that WSN model will be of low power entirely such that the battery will be modeled analytically as a reservoir of energy that is drained by a rate depended on the loads [26], [27]. All these will ensure that the energy needed for running the network, is reduced to minimum level.

VI. Conclusion And Future Work

The paper has considered various cases of pipelines breaks and independent technologies of monitoring oil pipelines, with advantages and disadvantages of each using each technology independently. Proposed integrated technologies were also looked into and a preferred proposed architectural combination of the technologies was fully discussed. It is believed that proper configuration of the technologies will go a long way to reduce or completely eliminate every form of loss along the pipeline and safeguard the environment from hazardous events. This is just an overview of the integrated networks for pipeline monitoring; the future work will be to consider all the protocols needed for this configuration to work effectively and also to see if the proposed technology is practically possible and financially viable.

References

[1] N. Mohamed and I. Jawhar, "A Fault Tolerant Wired/Wireless Sensor Network Architecture for Monitoring Pipeline Infrastructures" presented at the 2nd International Conference on Sensor Technologies and Applications (SENSORCOMM 2008), France, 2008.

[2] J. Frings, "Enhanced Pipeline Monitoring with Fiber Optics Sensors," presented at the 6th Pipeline Technology conference, Germany, 2011.

[3] P. Parfomak, "Pipeline Safety and Security: Federal Programs," D. o. H. Security, Ed., ed. United States of America, 2010, pp. 1-20.

[4] N. Mohamed, I. Jawhar, J. Al-Jaroodi, and Z. Liren, "Sensor Network Architectures for Monitoring Underwater Pipelines " *Sensors - Open Access Journal,* vol. Vol. 11, p. 27, 2011.

[5] How Pipeline Vandals Cripple Fuel Supply-NNPC....Incurs over N174 billion in Products losses and Pipeline Repairs. Available: http://www.nnpcgroup.com/PublicRelations/NNPCinthenews/tabid/92/articleType/ArticleView/articleId/68/How-Pipeline-Vandals-Cripple-Fuel-Supply--NNPCIncurs-over-N174-billion-in-products-losses-pipeline-repairs.aspx (accessed on 26 December 2012)

[6] N. Vatte and A. Sagar, "Real-Time Surveillance and Monitoring of Pipelines," Schlumberger Limited, Nigeria 2010.

[7] A. Diagnostics. Leak Detection for Pipelines, Tank Farms and Chemical Plants (LEOS). Available: http://www.areva-diagnostics.de/en/leak-detection.html (accessed on 29 December 2012)

[8] Leak Detection. Available: http://en.wikipedia.org/wiki/Leak_detection. (accessed on 25 December 2012).

[9] List of pipeline accident. Available: http://en.wikipedia.org/wiki/List_of_pipeline_accidents/ (accessed on 20 December 2012).

[10] I. Stoianov, L. Nachman, S. Madden, T. Tokmouline, and Acm, *PIPENET: A wireless sensor network for pipeline monitoring*, 2007.

[11] *Pipeline Guerrillas*. Available: http://www.offshore-technology.com/features/feature1165 (accessed on 16 December 2012)

[12] Oil spill in Dalian China. Available: http://www.boston.com/bigpicture/2010/07/oil_spill_in_dalian_china.html (accessed on 26 December 2012)

[13] I. Jawhar, N. Mohamed, and K. Shuaib, "A framework for pipeline infrastructure monitoring using wireless sensor networks," presented at the 2007 Wireless Telecommunications Symposium, United Arab Emirates, 2007.

[14] O. Aboderin, "Wireless Sensor Networks: Architecture, Applications and Future Development," *9th Research Seminar Series Workshop, School of Engineering, Design and Technology, University of Bradford, United Kingdom*, April 2010.

[15] E. Tapanes, "Fibre Optic Sensing Solutions for Real-Time Pipeline Integrity Monitoring," *The Australian Pipeliner,* vol. 4, p. 10, 1999.

[16] D. Inaudi, B. Glisic, A. Figini, and R. Walder, "Pipeline Leakage Dectection and Localization using Distributed Fiber Optic Sensing," presented at the Rio Pipeline Conference & Exposition 2007, Rio de Janeiro, Brazil, 2007.

[17] G. L. Burkhardt and A. E. Crouch, "Real Time Monitoring of Pipelines for Third-Party Contact," Sensor Systems and NDE Technology Department Applied Physics Division, Southwest Research Institute, Texas, USA 2003.

[18] AT&T Cables Vandalized, $250,000 Reward Offered for Information. Available: http://www.nbcsandiego.com/news/local/ATT-Cables-Cut-Vandalized-250000-Reward-Offered-159155295.html (accessed on 25 December 2012).

[19] S. SADOVNYCHIY, "Unmanned Aerial Vehicle System for Pipeline Inspection," Athens, Grecce, 2004. monitoring," *Ad Hoc Networks,* vol. 9, May 2011.

[20] G. Tuna, T. V. Mumcu, K. Gulez, V. C. Gungor, and H. Erturk, "Unmanned Aerial Vehicle-Aided Wireless Sensor Network Deployment System for Post-disaster Monitoring," in *Emerging Intelligent Computing Technology and Applications*. vol. 304, D. S. Huang, P. Gupta, X. Zhang, and P. Premaratne, Eds., ed: Springer-Verlag Berlin Heidelberg, 2012, pp. 298-305.

[21] C. Chen, Y. Tan, and L. Xing, "Study on Application of Unmanned Aerial Vehicle for Disaster Monitoring," *Research Journal of Chemistry and Environment,* vol. 16, Nov 2012.

[22] D. Manolakis, "Detection Algorithms for Hyperspectral Imaging Applications " Massachusetts Institute of Technology, Massachusetts p. 85, 2002.

[23] W. E. Roper and S. Dutta, "Oil Spill and Pipeline Condition Assessment Using Remote Sensing and Data Visualization Management Systems," George Mason University, United States of America2006.

[24] S. Kazem, M. Daniel, and Z. Taieb, *Wireless Sensor Networks: Technology, Protocol, and Applications*, 1st ed. New York: John Wiley and sons, 2007.

[25] J. Allen and B. Walsh, "Enhanced Oil Spill Surveillance, Detection and Monitoring Through the Applied Technology of Unmanned Air Systems" presented at the International Oil Spill Conference, Georgia, USA, 2008.

[26] Y. Guo, F. Kong, D. Zhu, A. s. Tosun, and Q. Deng, "Sensor Placement for Lifetime Maximization in Monitoring Oil Pipelines " presented at the First International Conference on Cyber-Physical Systems (ICCPS), Stockholm, Sweden, 2010.

[27] M. Aboelaze and F. Aloul, "Current and future trends in sensor networks: A survey," presented at the 2005 International Conference on Wireless and Optical Communications Networks, Dubai, UAE, 2005.

Trade-Off Between Paging and Tracking Area Update Procedures in LTE Networks

Syed Saqlain Ali
IT-Institute of Telecommunication, University of Aveiro, Aveiro Portugal
syedsaqlain@av.it.pt

***Abstract*—The enormous growth in the telecommunication market results in the optimised and cost effective networks for the service providers and the operators. In the study of Location Management (LM) of cellular networks, user cost-efficiency is one of the major challenges. Tracking Area (TA) manages and represents the location of User Equipment (UE). In general, small TA_s require more Tracking Area Update (TAU) but less paging procedures while large TA_s require less TAU_s but more paging procedures. A balance between Tracking Area Update (TAU) and Paging must exist in order to free the resources in the air interface. The aim of this work is to find out trade-off between signalling overheads in terms of TAU and the paging procedures. The Concept of Tracking Area List (TAL) management is used to overcome the problem of PingPong effect and frequent number of Tracking Area Update (TAU) procedures to minimize the signalling overheads. In the paper, two different TAL configuration have been used in order to show the trade-off between paging and TAU in terms of number of TA.**

***Index Terms*—LTE, Trade-off, Paging, Tracking Area Update (TAU), Tracking Areas(TAs), Tracking Area List (TAL)Management.**

I. Introduction

In Long Term Evolution (LTE) Tracking Area List (TAL) is a feature introduced in Third Generation Partnership Project (3GPP) Release 8. TAL management assigns more than one tracking area to UE or multiple tracking areas to a cell. All TA that are in TAL in which the UE is registered are served by the same serving Mobility Management Entity (MME). When the UE registers with the network the MME assigns a list of TA to the UE by making the centre of these set of TA close to UE current location, so the chance of initializing another TAU by the UE will be reduced. It has been seen that the TAL can reduce the frequent TAU when the UE keep moving between two or more cells in different TA_s which is called the Ping-Pong effect [1].

The rest of the paper is structured as follows: Section (II) provides the state of the art of proposed methods. In Section (III), research methods and techniques are described along with the demonstration on two scenarios 1 and 2, which shows the trade-off between paging and TAU in terms of number of Tracking Areas. Results are presented in section (IV) along with the comparison of scenario 1 and 2. Finally section (V) concludes with the main results.

II. STATE OF THE ART

Paging procedure is used to find the exact location of a user at cell level. Each attempt of determining the location of a user is referred to as a polling cycle. During each polling cycle, polling signals are sent over the downlink control channel to all the cells which are in TAL where a user is likely to be present. All the users in the TAL listen to the paging message and only the called user sends a response message back via uplink control channel. There is a timeout period in every cycle and if the user does not respond within that time frame than other group of cells in TAL will be paged for next time period [2]. The MME is responsible for keeping track of the user location while the UE is in EPS-Connection Management (ECM-IDLE). When there is a need to deliver downlink data to an ECM-IDLE UE, the MME sends a paging message to all the eNodeBs in its current TA as shown in left-side of Figure 1, and the eNodeBs page the UE over the radio interface. On receipt of a paging message, the UE performs a service request procedure which results in moving the UE to ECM-CONNECTED state.

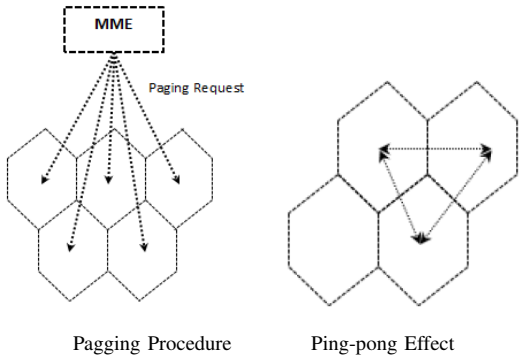


Fig. 1. Paging Procedure and Ping-pong Effect

A. Paging Frames and Paging Occasions

UE can be paged only in certain frames and sub-frames. Frames in which paging is allowed are called Paging Frames (PF) and the Sub-Frames are called Paging Occasions (PO). The Figure 2 below describes the concept PF and PO. In the paging procedure the value of nB is $1/16$ which gives the number of PO per radio frame equals to 0.0625 and the total no of PO per DRx cycle is equal to 8 [3].

B. Parameter nB

The capacity of EnodeB is determined by the number of available PO_s per radio frame. This parameter is used to find

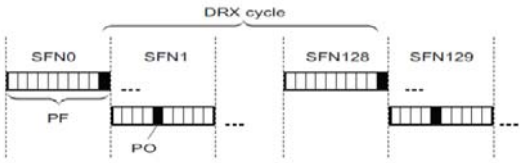


Fig. 2. Paging Frames and Paging Occasions

the frequency of PF and PO, nB can be $1/32T$, $1/16T$, $1/8T$, $1/4T$, $1/2T$, T, $2T$, $4T$, where T is the default Paging Cycle. The Figure 3 below illustrates the effect of nB on number of PF and PO [3].

- If nB >= T there will be nB/T POs per PF. All radio frames may be used for paging.
- If nB < T there will be one PO per PF.
- No of PO /Radio frame= nB/T

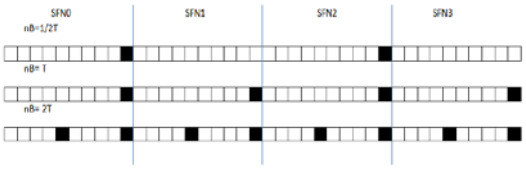


Fig. 3. nB setting on number of PF and PO

C. Tracking Area Update Procedures

To allow the network to contact an ECM-IDLE UE, the UE updates the network as to its new location whenever it moves out of its current Tracking Area (TA), this procedure is called Tracking Area Update [4].

D. Ping-Pong Effect

UE_s can move back and forth at the borders of TA_s between two or more TA_s. This effect is called Ping-Pong effect as shown in the right side of Figure 1. This effect causes excessive TAU signalling overhead. By reducing this effect the TAU can be reduced and this can be done by introducing the concept of TAL [5].

E. Multiple TA_s

In this scheme cell belongs to only one TA and UE can be assigned by multiple TA using the list of TA_s. UE does not perform TAU while crossing the boundaries between the assigned lists of TA. In this scheme TA_s are non-overlapping. By introducing the concept of TAL provides more flexibility to the network operator in TA dimensioning and planning. The other schemes results in the reduction of signalling overhead due to TAU and results in increasing the paging overhead but the paging overhead considered as less critical issue as compared to TAU. The concept of multiple TA is more preferred scheme as compare to overlapping TA concept [6] [7].

F. Requirement for Tracking Area Concept

The concept of TAL is similar to the Routing area concepts in Universal Terrestrial Radio Access Network/GSM-EDGE Radio Access Network (UTRAN/GERAN). Each cell broadcasts one Tracking Area Identity (TA-ID). When the UE in idle state enters a new TA which is not assigned, it will perform TAU procedure. So the UE will be assigned to the new TA plus an additional TA as a result of TAU procedure. These additional TA_s are treated in the same way as in the single TA case, meaning that as long as the terminal moves within the TA_s it has been assigned to it will not perform any TAU procedures except periodical updates [8].

1) Flexibility: In configuring the TA_s in LTE networks there always exists a trade-off between TA size and paging load:

- TA_s with small size leads to less paging load but they generate more TAU messages.
- TA_s with large size leads to more paging load but they generate less TAU messages.

It may be useful to have a TA concept that allows the flexibility on the size of TA a UE will be paged in. Like stationary UE can be assigned to small area while moving UE can be assigned to a large paging area as proposed in (distance based paging, velocity based paging) [9].

G. Example of TAL assignment

A scenario shown in Figure 4 describes the assignment of TAL to the UE_s. In this example TAL_1 contains TA_1, TA_2, TA_4, TA_5 and TAL_2 contains TA_2, TA_7, TA_9 and TA_10. The network assigns TAL_1 to UE_1 and TAL_2 to UE_2. UE_1 will only need TAU while moving from TA_2 to TA_7 because UE_1 do not have TA_7 in its TAL while it will not need any TAU when moving to the assigned TA in TAL. UE_2 will not need TAU while moving from TA_2 to TA_7 it will need TAU to the other TA which is not in the TAL of UE_2 [10].

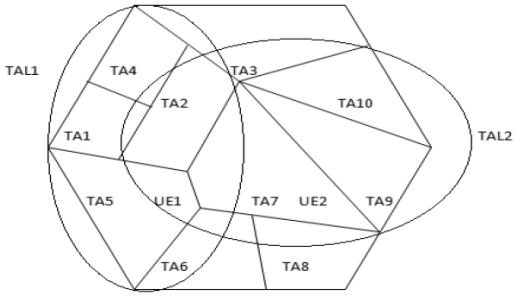


Fig. 4. TAL Assignment Example

III. Methods and Techniques

By considering a large city in which there are $800,000$ users serving by a network in which users are moving (changing TA_s) during the office starting and ending hours. As there is a large number of users that are moving and coming back towards residential and office area it will require an efficient assignment of TA_s for the UE_s. The scenarios 1 and 2 follow

TABLE I
TAU LOAD:SCENARIO 1

Number of Subscribers	Periodic TAU	Leave TAL_1	Leave TAL_2	Total TAU
560000	$5.6x10^5$	$5.6x10^5$	$5.6x10^5$	$1.68x10^6$
240000	$2.4x10^5$	-	-	$2.4x10^5$
Total TAU Load	$8x10^5$	$5.6x10^5$	$8x10^5$	$1.92x10^6$

the assumptions made in this section. The Total number of BHCA (including VoIP, Video, Web Browsing and E-mail) for $800,000$ Users can be calculated as follows:

Total Number of BHCA = Number of Subscriber x BHCA Per subscriber.

Total Number of BHCA for all UEs = $800,000 \times 1.15 = 9.2 \times 10^5$.

A. Planing Tracking Areas

The process of planning of TA includes: determining the TA borders and the configuration of the list of TA_s so that the areas with excessive TAU signalling will be avoided [11] [12] . In planning of TA the following rules apply:

- Areas with frequent TAU signalling must be located in low traffic area which results for eNodeB to cope up with extra signalling caused by TAU.
- In order to minimize the TAU signalling TA and TAL must be planned well. This can be done by considering the movement of users in the network like busy roads, public places, railway line etc. It must cross few TAL borders as possible.

B. Scenario 1:TAL Management

By using the assumption made above(see Section III). In this TA configuration we assume that the UE_s are assigned small number of TA as compared to the TA in scenario-2. Figure 5 shows the TAL_1 and TAL_2 configuration in which UE_s are registered with TAL_1 and TAL_2. TAL_1 consists of TA_1, TA_2, TA_3, TA_4, TA_5, TA_6, TA_7, TA_8, and TA_9, whereas TAL_2 contains TA_9, TA_10, TA_11 and TA_12. For the UE_s registered with TAL_1 they will need only Periodic TAU and they will not need any extra TAU as long as they will move inside the TAL_1. In this scenario 70 percent of the total users are moving in the network while 30 percent are the stationary users and the periodic timer will initiate TAU after every 24 hours. For the 30 percent of UE_s only periodic TAU is needed while for the 70 percent of UE_s they will need periodic TAU as well as TAU whenever they enters the new TA during the office starting hours and ending hours there will be two peaks on the graph as shown in the Figure 8. Tables I, II and III shows the calculation of TAU load, Paging load and Total signalling overhead for scenario-1 during 24 hours a day [13] [14].

C. TA Design Optimisation

As seen in the above scenario 1 most number of the users change their TA during the offices hours which creates an excessive TAU signalling overhead. As it has been seen that having TA_s of very small size (e.g., one cell per TA) virtually eliminates paging, but causes excessive TAU, whereas large number of tracking areas reduces the TAU overhead but it will increase the paging overhead. So the main objective in TA planning is to reach a balance between TAU and Paging. Scenario 2 gives an example with large number of TA and by observing the results of two scenarios TA design optimisation can be done. Figure 6 illustrate the trade-off between the two parameters.

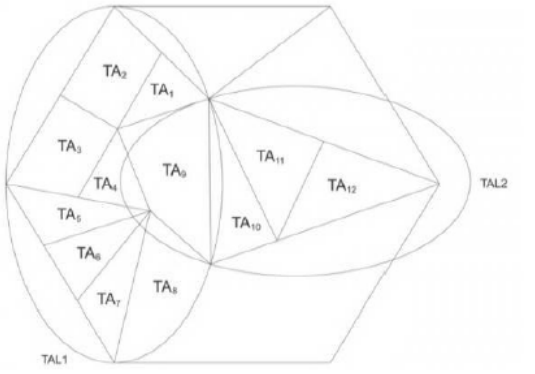


Fig. 5. Scenario 1:TAL Management

TABLE II
PAGING LOAD: SCENARIO 1

UE_s	**800000**
TA_s in TAL	9
eNodeB per TAL	25
Total paging load in TAL	800000 x 1.15 x 25 = $2.30x10^7$

TABLE III
SIGNALLING OVERHEADS SCENARIO 1

Signalling Overhead	Total Paging and TAU Load
Paging	$2.30x10^7$
TAU	$1.92x10^6$
Total Signalling Overhead	$2.49x10^7$

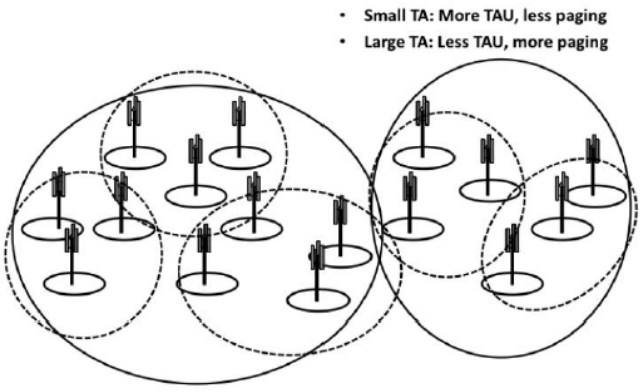


Fig. 6. Illustration of TAU and Paging trade-off

TABLE IV
TAU LOAD:SCENARIO 1

Number of Subscribers	Periodic TAU	Leave TAL_1	Leave TAL_2	Total TAU
400000	$4.0x10^5$	$4.0x10^5$	$4.0x10^5$	$1.20x10^6$
400000	$4.0x10^5$	-	-	$4.0x10^5$
Total TAU Load	$8.0x10^5$	$4.0x10^5$	$4.0x10^5$	$1.60x10^6$

TABLE V
PAGING LOAD: SCENARIO 2

UE_s	**800000**
TA_s in TAL	11
eNodeB per TAL	30
Total paging load in TAL	800000 x 1.15 x 30 = $2.76x10^7$

D. Scenario 2: TAL Management

In this TA configuration we assume that the UE_s are assigned large number of TA as compared to the TA in scenario-1. (see section III-B). Figure 7 shows the TAL_1 and TAL_2 configuration in which UE_s are registered with TAL_1 and TAL_2. In this scenario 50 percent of the total users are moving in the network while 50 percent are the stationary users and the periodic tracking update timer will initiate TAU after every 24 hours. For the 50 percent of UE_s only periodic TAU is needed while for the 50 percent of UE_s they will need periodic TAU as well as TAU whenever they enters the new tracking area during the office starting hours and ending hours there will be two peaks on the graph as shown in the Figure 8. Table IV, V and VI shows the calculation of TAU load, Paging load and Total signalling overhead for scenario-2 during 24 hours a day [13,14].

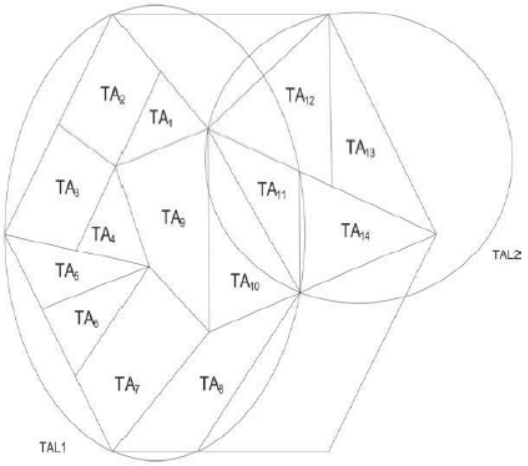


Fig. 7. Scenario 2: TAL Management

IV. RESULTS

Tables III and VI show the signalling overhead in terms of paging and TAU load for scenario 1 and 2. As shown in the Figure 8, scenario 1 generates higher signalling overhead in terms of TAU and generates less paging overhead in terms of paging load as the number of TA in TAL is small as compare

TABLE VI
SIGNALLING OVERHEADS SCENARIO 2

Signalling Overhead	Total Paging and TAU Load
Paging	$2.76x10^7$
TAU	$1.60x10^6$
Total Signalling Overhead	$2.92x10^7$

to the number of TAs in scenario 2, where the signalling overhead in terms of TAU is less than the scenario 1 but it generates higher signalling overhead in terms of paging due to large number of TA_s in TAL as MME has to send the paging request to all the TA_s in which user is registered with. But as the size of TAU message is very large as compare to the size of paging request message, TAU will occupy much air resources as compare to paging request message. So reducing the signalling overhead due to TAU is much important in order to use less resources in the radio interface.

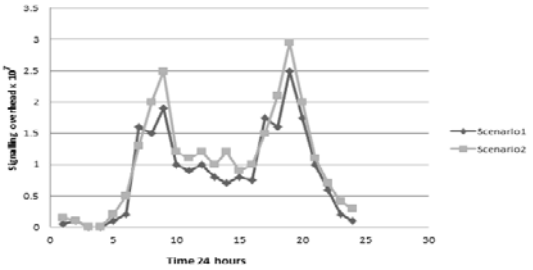


Fig. 8. Comparison of Scenario 1 and 2

V. CONCLUSION

The paper provides the benefit of TA_s planning and optimisation in order to reduce the signalling overhead for improving the performance of cellular networks. The above scenarios provide the tracking area list (TAL) management in terms of paging and TAU signalling overhead. Figure 9 shows a trade-off between paging and TAU procedures: As the size of TA increases the paging overhead tends to increase and TAU will decrease and if the size of TA decreases the paging overhead tends to decrease as TAU will tend to increase. At some point of the configuration between number of TA_s and signalling overhead a trade-off exists and at this point there will be balance between paging and TAU which optimises the signalling overhead in the network. Also approaches like UE assistance and self Organization can be future line of work.

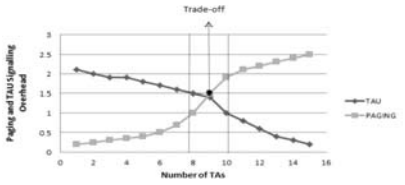


Fig. 9. Trade-off between Paging and TAU

REFERENCES

[1] "General packet radio service (gprs) enhancements for e-utran access," 3GPP TS 23.401 V.9.4 .0, Release 9.

[2] "Ericsson enhanced paging procedures," 3GPP TSG RAN WG2,57.

[3] S. A. V.Srinivasa Rao and L. E. Rambabu Gajula, "Signaling procedures in lte," webbuyersguide.com, March 12 2010.

[4] "General packet radio service (gprs) enhancements for e-utran access," 3GPP TS 23.401 V.9.4 .0, Release 9.

[5] "Nokia ping-pong control," 3GPP TSG-RAN-WG2 Meeting 57. R2-070713.

[6] "Samsung: Use of tracking area- and cell identity," 3GPP TSG-RAN WG2 Meeting 57, R2-070680,.

[7] "Qualcomm europe performance criteria for tracking area concepts." 3GPP TSG-RAN WG2-57 R2-070718.

[8] "Ericsson enhanced tracking area concept in sae/lte," 3GPP TSG RAN WG3-53, 28 August 1 September, 2006.

[9] "Nec, tracking area concepts and influence on paging load," Technical Report in 3GPP TSG RAN2 Meeting, R2-070655, 2007.

[10] S. M. Razavi and D. Yuan, "Performance improvement of lte tracking area design: a re-optimisation approach," in *Proc. of the 6th ACM International Workshop on Mobility Management and Wireless Access (MobiWac 08)*. ACM, 2008, pp. 77–84.

[11] "Motorola lte rf planning guide version: 1.2."

[12] "Lte-ta-planning," http://www.telecom-cloud.net/wp-content/uploads/2010/09/LTE-TA-Planning.pdf.

[13] F. G. S.Modarres Razavi, D. Yuan and J.Moe, ", dynamic tracking area list configuration and performance evaluation in lte," in *Proc. of IEEE Global Communications Conference Workshop (GLOBECOM Workshop 10)*. IEEE, 2010.

[14] F. G. S. Modarres Razavi, D. Yuan and J. Moe, "Exploiting tracking area list for improving signaling overhead in lte," in *in Proc. of IEEE Vehicular Technology Conference*, 2010.

SESSION 4

MICROWAVE AND CIRCUIT DESIGN

Q-Band Short-Slot Hybrid Coupler in Gap Waveguide

Bilal Hussain
Faculty of Engineering University of Porto (FEUP)
Porto, Portugal
bhussain@inescporto.pt

Iman Kianpour
Faculty of Engineering University of Porto (FEUP)
Porto, Portugal
ikian@inescporto.pt

***Abstract*— This paper presents a Q-Band 3 dB short-slot hybrid coupler designed using recently evolved gap waveguide technology. The coupler structure is manufactured with an allowable gap between two metal blocks, in such a way that there is neither requirement to electrical contact nor alignment between the blocks. This is a major manufacturing advantage compared to normal rectangular waveguide. This coupler works from 37.5-39.5 GHz (5% bandwidth) with an amplitude variation of ±0.25 dB and provides a phase of 90˚ at the coupled port. Also the isolation is found to be better than -20 dB between input and isolated port. The design technique used is similar to rectangular waveguides but with slight structural modifications of Gap waveguides.**

***Index Terms*—3 dB coupler, Groove Gap waveguide, Riblet coupler**

I. INTRODUCTION

COUPLERS are extensively used in microwave circuits. Their applications include: provision of signal sample for monitoring and measurement purposes, combining feed to and from antennas, antenna beam forming, feedback, splitting different signals, etc. Waveguide couplers are mostly used for high power and antenna applications. In waveguides, couplers are designed by providing coupling slots in the narrow or broad wall of the guide, known as short-slot or bethe-hole coupler respectively. In the traditional rectangular waveguides, it is difficult to manufacture components for frequencies above 60 GHz. For conventional rectangular waveguides, structure needs to have electrical contacts between broad and narrow walls. When interfaces such as mechanical bends are introduced, it becomes mechanically challenging to provide good electrical contact. To some extent the problem is solved by dividing the structure in to two halves, and later on joining them by screws. But this topology is mainly challenged by the fact that at high frequencies the dimensions get really small, and it becomes almost impossible to provide good electrical contact without any field leakage. The problems of good electrical contact and alignment between the two halves are solved by introducing gap waveguide technique. This newly developed gap waveguide technology is well described in [1]-[6]. Between two parallel plates, a special quasi-periodic structure is designed, as to guide the incoming wave along the desired path and creating cutoff for transmission in any other direction. The special structure [1]-[2] can e.g. be composed of square pins with certain period and dimension of the pins but other structures are also possible [5].. This structure mimics the ideal PMC (perfect magnetic conductor) [8], thus the wave cannot propagate along this surface when a top metal plate is placed at a distance smaller than λ/4. Using this altered structure, two types of waveguiding techniques are introduced i.e., ridge gap waveguide [1]-[2] and groove gap waveguide [4]. Ridge gap waveguide resembles more like a microstrip, while groove gap waveguide can be compared to rectangular waveguide. Fig. 2 shows a typical groove gap waveguide structure. In [4], it was proven that groove gap waveguide is having certain similarities with rectangular waveguides. The desired mode in groove gap waveguide has field distribution similar to that of the dominant mode of rectangular waveguide TE_{10}. The guided wavelength for desired mode is also quite equal in both structures. Although, the field equations for groove gap waveguides are not established yet, but keeping in mind these similarities, design techniques of rectangular waveguides can be applied to groove gap waveguides.

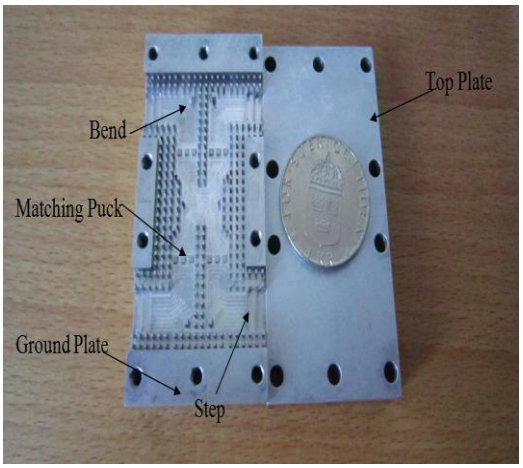


Fig. 1 Manufactured Coupler in Groove Gap waveguide with WR-28 ports

This paper is organized as follows: Design methodology is defined the next section. Also the technological bottle necks and their solutions are also presented. Section III presents measurements and the techniques used to obtain them. Section IV provides conclusion and comparison with the existing technologies such as SIW. Also, the mechanical design problems and resulting insertion loss is also discussed briefly in this section.

II. SHORT-SLOT COUPLER

Short-slot or Riblet coupler was first introduced by H.J Riblet in 1947. It's a rectangular waveguide coupler, constructed by providing a coupling section along the common narrow wall of two waveguides. The theory of operation is presented in [7]. As mentioned earlier that the groove gap waveguide shares similarities with rectangular waveguides. Therefore, a 3 dB coupler is designed in groove gap technology using the design rules of rectangular waveguides.

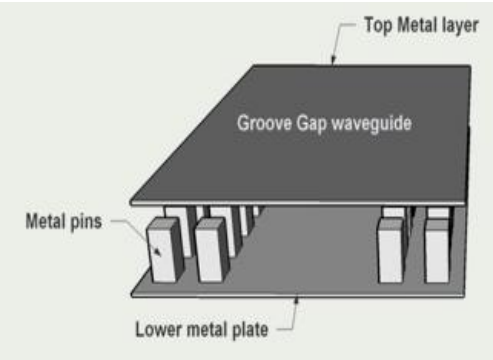


Fig. 2 Groove Gap Waveguide (taken from [1])

Based on [4], a structure having two groove gap waveguides as shown in Fig. 3(a), in Q-band (30-40 GHz) was first simulated using HFSS to check for isolation between two guides. Both groove gap waveguides were separated by two rows of pins. These pins were designed using dimensions from [4], and frequency scaling them for the desired band. It was found that both groove gap waveguides have isolation better than -40 dB and return loss was found to be below -30 dB over the entire band. Before designing the coupler, we needed to analyze the structure of gap waveguides. Gap waveguides can be divided in to two structures: Bed of nails and Lid. If we compare it with rectangular waveguide, we can see that voltage loop is easy to realize in this structure, as voltage is contained between top and bottom plate. But when we tried to realize current loop, we found out that there is no continuous path between side walls. There is air present between side wall and top lid. This discontinuity will make the gauss law discontinuous at this point, therefore, it makes impossible to use Maxwell equations. Therefore, approach used here is to simulate the groove gap waveguide structure first, and then observe the behavior of TE_{10} mode in the structure. This provided basis for using rectangular waveguide equations for designing a coupler in groove gap waveguides, as the behavior of dominant mode is similar as in conventional waveguides. To design a 3 dB short-slot coupler, a coupling slot between the two waveguides was provided by removing pins from the central 2 rows. The length of this slot is 'L' as shown in the Fig. 3(b). Initially, slot length was chosen greater than half guided wavelength, where guided wavelength is calculated using (1).

$$\lambda_g=\lambda_0/\left(1-\lambda_0^2/\lambda_c^2\right)^{1/2} \qquad (1)$$

λ_g is the guided wavelength, λ_0 is the free space wavelength and λ_c is the cutoff wavelength of rectangular waveguide having broad dimension same as groove gap waveguide mentioned in Fig. 3(a).

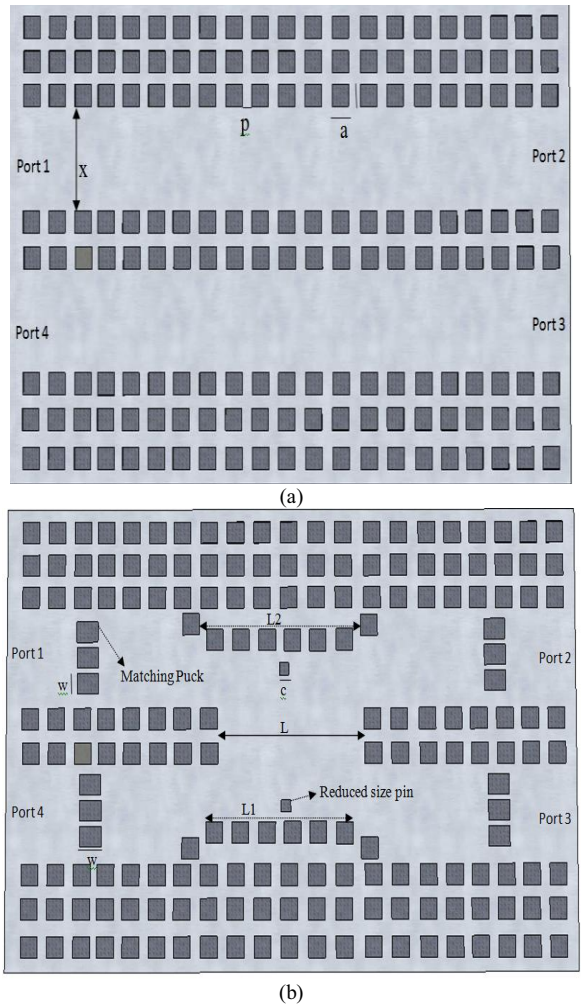


Fig. 3(a) Top view of two groove gap waveguides lying side by side. a=0.62mm, x=6.53mm, p=1.03mm, Height of the Pin=2.06mm, Air Gap=0.43mm. (b) Top view of coupler in groove gap waveguide. L=7.9mm, L1=8.9mm, L2=10.5mm, w=1.02mm, c=0.38mm, Height of Puck=0.48mm.

Later on this length was adjusted to meet the bandwidth requirement. As established in [7], in the coupling region of 'L', width of the waveguide is doubled. To have 3 dB coupling, a TE_{20} mode is desired. But the doubled width of this coupling section can excite TE_{30} mode also. To stop higher order mode excitation, width of the coupling section is adjusted by providing 'L1' and 'L2' sections. These sections are longer than the coupling slot and are displaced to match the width required for the propagation of TE_{20} mode. A good approximation for the width can be obtained by solving the cut-off frequency equation of rectangular waveguides for TE_{20} mode. Yet to obtain a wider bandwidth, an additional section is added by providing two shorter pins (reduced size pins) as shown in Fig. 3(b). Bandwidth can be increased by providing a pin with increased size in the middle of the coupling section, but matching is difficult with such an arrangement.

Introducing the slot 'L' and width adjustments 'L1' and 'L2' creates reactive impedance, as in the case of rectangular

waveguides [10]. In rectangular waveguides, matching is achieved by introducing tuning screws or posts along the waveguide. Similar to this, matching pucks are provided in the groove gap waveguide structure. In fact these pucks are pins with reduced height and increased size. Such a discontinuity is applicable for modes with no field variation in y-direction. Position of these pucks was found using parametric sweep of HFSS. It is interesting to note that matching can be achieved by using one puck only. But to provide a wider bandwidth, three pucks provide optimal performance. In this proposed design, all the pucks are of the same height and size, but a variation in the dimensions can provide increased bandwidth.

To interface groove gap waveguide coupler with standard rectangular waveguide flange, 90˚ H-plane bends were introduced. Since the structure was never used before in order to interface with rectangular waveguides. Introduction of bends was challenging as well as required an in depth study of bends in rectangular waveguides. In conventional waveguides a bend always increases the width of the guide at the corner in H-Plane. This increase in width causes the propagation of higher order modes, thus introduces the loss in waveguide. To overcome this loss, width of this corner is adjusted such that it allows only the required mode to propagate. To adjust the width, waveguides are cut at an angle. This type of cut is known as x-y cut [9]. Similarly, a pin with the same height as other pins but with increased size is introduced. This pin mimics the conventional x-y cut of rectangular waveguides. H-plane bend in groove gap waveguide is shown in Fig. 4. Moreover, this bend also matched the broad dimensions of groove gap waveguide to standard rectangular waveguide flange. The narrow dimensions of groove gap waveguide are increased by providing quarter wavelength steps in the ground plate as indicated in Fig. 1. Simulation results for groove gap coupler are shown in Fig. 4. These results were obtained after introducing afore mentioned bends and steps in the ground plate.

III. MEASUREMENT RESULTS

Measurements were performed with a PNA (Power Network Analyzer) from Agilent (E-8363B series). During the measurements, TRL (Through-Reflect-Line) calibration was done on the WR-28 rectangular waveguide ports. Fig. 6(a) shows the measured S-parameters for groove gap waveguide coupler. These results show good agreement with the simulations of Fig. 5. It is interesting to note that return loss and isolation for both measured and simulated results are identical. Moreover, the phase difference over the entire range is 90˚ with a phase error of ±2.5˚ fig. 6(b). Also, these measurements were repeated with no screw attached to ground plate and top lid, yet the results did not change. Ideally, screws are not required but to avoid misalignment between top and bottom plate, two screws can be used at the opposite corner of the structure. This establishes the fact that ground plate and top lid need not to have an electrical contact. It also can be argued that the rectangular waveguide port can provide a ground contact for the structure. But the thickness of the metal is enough to provide severe skin depth effect at these frequencies. Moreover, the length of the waveguide is nine times the quarter wavelength, any contact provided at the input, will be lost within first few centimeters. Also the quarter wavelength steps provide an efficient transition from rectangular waveguide to groove gap waveguide. Therefore, we can say that measurements were quite reliable and repeatable in nature.

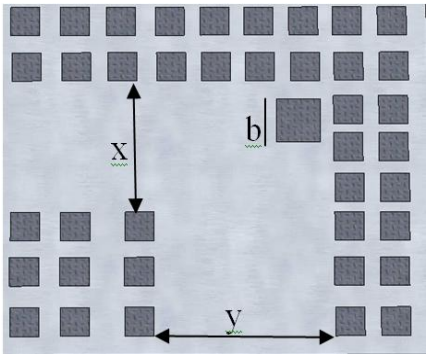


Fig. 4 Top view of H-plane bend in Groove Gap waveguide x=6.53mm,y=7.11mm,b=0.8mm

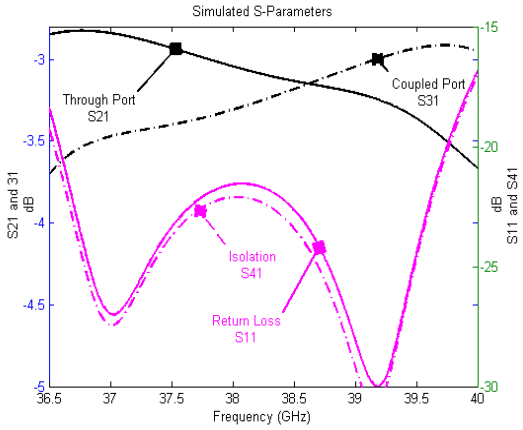


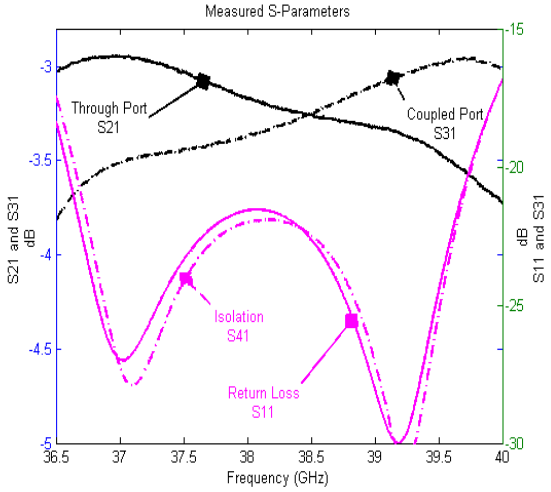
Fig. 5 Simulated S-Parameters

IV. CONCLUSIONS

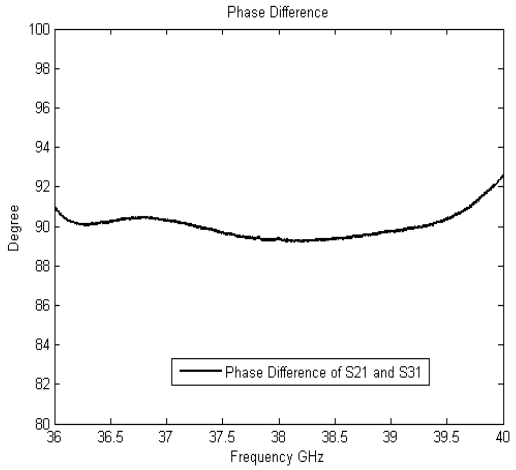
In this paper newly introduced groove gap waveguide technology is used to design a 3 dB coupler with a fractional bandwidth of 5%. Good electrical performance is achieved with this new type of coupler. Thus groove gap waveguides can be used to manufacture microwave components without strict requirement on the metal contact between the manufactured blocks, especially at high frequencies where dimensions are small and very good surface finishing as well as manufacturing tolerance is needed to have good metal contact. The above designed coupler clearly provides manufacturing ease as compared to rectangular waveguide coupler at the same frequency. Moreover, dimensions of pins are not so tight, as required by SIW (Substrate Integrated Waveguide) [11]. The reason for this flexibility is the use of air instead of a substrate as in the case of SIW. Use of air also reduces the insertion loss of the guide, as dielectric loss tends to increase with increase in frequency. Another important feature of this design is that it uses conventional rectangular

waveguide flange, thus no transition is required. This also contributes to a low loss design as compared to SIW.

Moreover, insertion loss can be noticed in the measured S parameters. This loss is due to the manufacturing tolerance of the structure. As the machining process is used to manufacture bed of nails, some pins were misaligned in order to provide ease for mechanical Saw and reducing the manufacturing time significantly for an insignificant loss. With a more controlled process of manufacturing, performance can be improved further. Also there is a need to investigate techniques such that it provides same electrical performance, with a more mechanically feasible design. Especially the sections L1 and L2 need to be simpler for a fast machining process. Also mechanical casting techniques are applicable for the manufacturing of bed of nails.



(a)



(b)

Fig. 6. (a) Measured S-Parameters (b) Phase Difference between coupled and through ports

REFERENCES

[1] P.-S. Kildal, A. Uz Zaman, E. Rajo-Iglesias, E. Alfonso and A. Valero-Nogueira, "Design and experimental verification of ridge gap waveguides in bed of nails for parallel plate mode suppression", *IET Microwaves, Antennas & Propagation*, Vol.5, Iss.3, pp. 262-270, March 2011.

[2] P.-S. Kildal, E. Alfonso, A. Valero-Nogueira, E. Rajo-Iglesias, "Local metamaterial-based waveguides in gaps between parallel metal plates", *IEEE Antennas and Wireless Propagation letters* ,Vol. 8, pp. 84-87, 2009.

[3] A. Valero-Nogueira, E. Alfonso, J. I. Herranz, P.-S. Kildal, "Experimental demonstration of local quasi-TEM gap modes in single-hard-wall waveguides", *IEEE Microwave and Wireless Components Letters*, Vol. 19, No.9, pp. 536-538, 2009.

[4] Rajo-Iglesias, P.-S Kildal,"Groove gap waveguide: A rectangular waveguide between contactless metal plates enabled by parallel-plate cut-off" in 4th European Conference on Antennas and Propagation (EucCAP 2010), 12-16 April 2010, Piscataway, NJ, USA, 2010,

[5] E. Rajo-Iglesias, P.-S. Kildal, "Numerical studies of bandwidth of parallel plate cut-off realized by bed of nails, corrugations and mushroom-type EBG for use in gap waveguides", IET Microwaves, Antennas & Propagation, Vol. 5, No 3, pp. 282-289, March 2011.

[6] E. Rajo-Iglesias, A. Uz Zaman, P.-S. Kildal, "Parallel plate cavity mode suppression in microstrip circuit packages using a lid of nails", IEEE Microwave and Wireless Components Letters, Vol. 20, No. 1, pp. 31-33, Dec. 2009.

[7] H.J.Riblet," A mathematical theory of Directional Coupler", I.R.E Proc, Vol.33, pp.1307-1313, 1947

[8] P.-S Kildal, "Definition of artificially soft and hard surfaces for electromagnetic waves" Electron. Lett, Vol.24, pp.168-170, 1998.

[9] Zhewang Ma, Taku Yamane, Eikichi Yamashita, "Analysis and design of H-Plane waveguide bends with compact size". University of Electro-Communication, Chofu-Shi Tokyo, Japan.

[10] N.Marcuvitz. Waveguide Han dBook. Vol 10 of MIT Rad Lab Series. McGraw-Hill. N.Y. 1948

[11] JinXin.Chen,Wei.Hong,Zhang.Cheng.Hao,Hao.Li,Ke.Wu,"Development of a Low Cost Microwave Mixer Using a Broad-band Substrate Integrated Waveguide (SIW) Coupler", IEEE Microwave and Wireless Components Letters, Vol. 20, No. 1, pp. 31-33, Dec. 2009.

An Ultra-Low Power Flash ADC for RFID and Wireless Sensing Applications

Iman Kianpour, Bilal Hussain, Jose Quevedo
Faculty of Engineering, University of Porto (FEUP)
Porto, Portugal
kiman@inescporto.pt, bhussain@inescporto.pt, jrquevedo85@gmail.com

***Abstract*—In this paper an ultra-low power Flash Analog to Digital Convertor (ADC) for Radio Frequency Identification (RFID) and Wireless Sensor Network (WSN) applications, is presented. Some techniques are used to reduce the power consumption and relatively elevate the speed of the ADC as much as possible. These techniques include a low power Track-and-Latch comparator with no static current, large resistors in the resistor string, an optimum encoder with only 2 stages using low-power design with the aid of low supply voltage of 0.7v for resistor string and 0.5v for all logic blocks. In this ADC, the occupied area is roughly equal to the area of 16 resistors in the string excluding decoupling caps and pads. The circuit designed in 0.18µm CMOS technology and post layout simulations show that the 4-bit ADC consumes almost 18µW at 36MS/s and 110nW at 0.1MS/s; however, it minimally dissipates only 31fJ per each conversion step. The results show that the proposed ADC could seriously compete with other low power ADCs in RFID sensing applications such as SAR ADCs.**

Keywords—Flash ADC; low-power design; RFID

I. INTRODUCTION

RFID is one of the modern applications of RF technology. In general, RFID is an Identification system like other techniques such as classic barcodes and biometrics. Usage of RFID is growing intensively these days, as you can see it in supermarkets for cashless payment, warehousing, asset management and also airport baggage control, library management, medical monitoring and access/security control, [1-3].

Each RFID system have two major components called transponder (tag), usually mounted on a label, and interrogator (reader). Tags usually contain information that should be read and interrogators are the equipment that read tag's information [1]. Tags consist of a chip and an antenna. The chip has data to send and the antenna is for sending which can be internal or external. Tags can be read-only or read-write. In general, there are two kinds of tags, active and passive tags [1]. Active tags take energy from battery and give them the ability of long range transmission and performing more sophisticated processes; but, at the cost of being larger, heavier and having limited lifetime [4]. Passive tags take energy from electromagnetic waves emitted by the reader, so they do not need battery and can be very small and very cheap and have unlimited lifetime, but, at the cost of shorter transmission range [2].

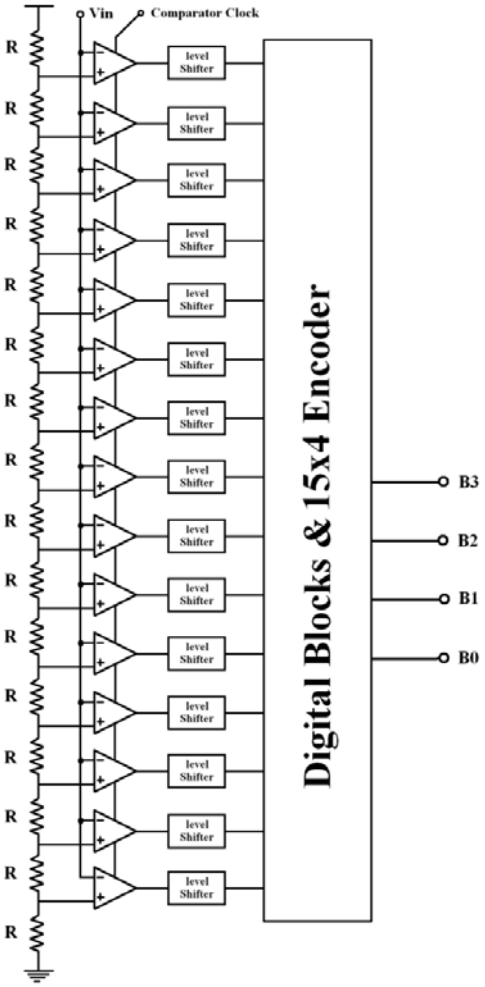


Fig. 1: Architecture of the presented ADC

This drawback comes from the fact that the electromagnetic waves emitted by the reader do not have abundant energy; therefore, the most necessary design consideration should be low power consumption in all building blocks. Additionally, a sensor can be integrated in a tag to sense ambient physical parameter; these tags usually called RFID Sensor Tags or Sensor-Embedded RFID (SE-RFID) [3] and also, Wireless

Identification and Sensing Platform (WISP) [5] which imply that these tags have both sensation and identification ability. Each sensor tag has 4 main components. 1) A sensor to sense. 2) Sensor interface to process output data of the sensor. 3) Analog RF front-end to modulate data and 4) an Antenna to communicate. Digital processing makes systems more capable, and gives them the ability of performing more sophisticated processes and can mitigate the challenging security problem of the tags. Therefore integrating an ADC and digital processor in architecture of RFID sensor tags is inevitable.

Among different types of ADCs, SAR ADC has been mostly used for moderate-speed, moderate-resolution applications [6]. Additionally, flash ADCs also could be utilized in design of power scavenging systems although concentration of designers is on charge redistribution SAR ADCs. Flash ADC has known as high-speed, high-power and low-resolution one; thus, usually prevented in low power applications.

This paper claims that flash ADCs also could be used for low-resolution, low-power systems, and would be a serious competitor for conventional low power SAR ADCs. Flash ADCs contain several building blocks, such as S/H, resistor ladder, comparator, encoder and latch. Among them S/H and latches are non-essential blocks and can be removed if the A/D converts dynamically [7].

II. Architecture of the Proposed ADC

Originally, this paper uses the traditional flash ADC architecture proposed in [6], but the difference is level shifters used due to several supply voltages. Architecture of the presented ADC is illustrated in Fig. 1. The input signal in a flash converter is fed to all comparators in parallel. Each comparator is also connected to a different node of a resistor string. Any comparator experiencing larger voltage (at the resistor string node) than Vin will have a 1 output while those connected to nodes with less voltage than Vin will have 0 outputs. Such an output code word is commonly referred to as a thermometer code since it looks quite similar to the mercury bar in a thermometer [6]. A 15 to 4 decoder is used to translate different voltages of the 16-resistor string into digital codes. In Fig. 1 all digital blocks and the encoder are integrated in one block just for simplicity. Digital blocks mean some digital circuits used for detecting the transition level from 1 to 0, and bubble error correction. They are illustrated in Fig. 2 in an example of a 3-bit Flash ADC. The NAND gate that has a 0 input connected to its inverting input and a 1 input connected to its non-inverting input detects the transition of the comparator outputs from 1s to 0s, and will have a 0 output. All other NAND-gate outputs will be 1, resulting in simpler encoding [6]. It also allows for bubble error correction with preventing more than one 0s at the output thermometer code. By using 3-input NANDs with that extra one connected to the output of the upper comparators, it can be guaranteed that no bubble error will occur for all possible input voltages.

In proposed design, 3 different voltages are used as power supplies which include 0.7v for resistor string, 1.2 for comparator and 0.5v for digital blocks and encoder. 0.7v supply of resistor string means that the input range can be 0 to 0.7v. Level shifters are only two consecutive NOT gates, the 1st one with 1.2v supply voltage of comparator and the 2nd one has the 0.5v power supply of digital blocks including encoder and bubble error correction. Such a high power supply voltage (1.2v) is commonly used in some RFID applications [8].

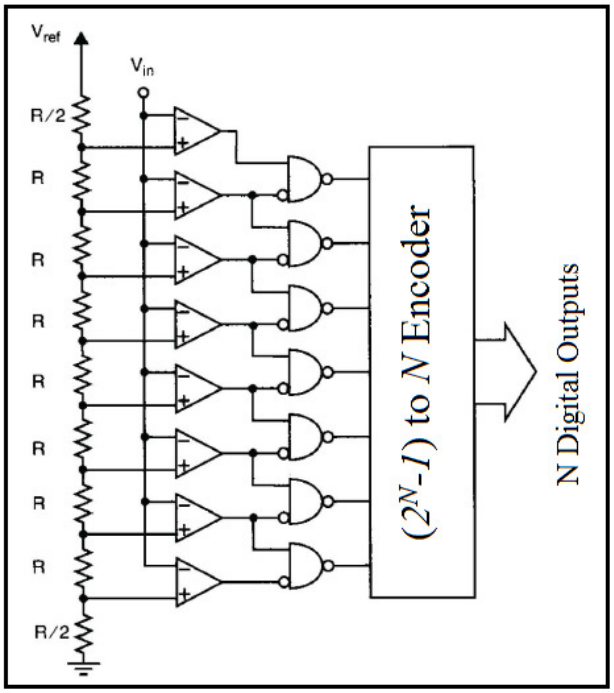


Fig. 2: Transition detection and bubble error correction for a typical 8-bit Flash A/D Convertor.

III. Comparator Design

Flash ADCs are fast but they require a large number of comparators, which typically take up a large area and are very power hungry. They need 2N-1 comparators, where N is the number of bits. So, the comparator is the most critical block in the architecture of Flash ADCs in terms of power dissipation. Thereby, in low power designs, Flash ADCs are usually being avoided; however, by profiting from a Track-and-Latch (T/L) comparator with no static current, a low-power Flash ADC could be achievable.

Track-and-Latch comparators in data convertors have advantages in comparison with continuous time (CT) comparators. T/L comparators have less noise sensitivity (especially kick-back noise) and high precision (a few µv). In this paper the comparator scheme is inspired from the one presented in [9], because it has no static current and therefore low power consumption is obtainable. Also, it could be compatible with ADC presented in this paper. In [9] the input differential pair was NMOS and they are changed to PMOS,

because the lowest output voltage of DAC, applied to the differential pair, is 0 and it has no capability to bias the input differential pair and the tail MOSFET. With this change the limitation transfers to the upper level of voltage. So, as can be inferred from Fig. 3, the upper limitation on input of the differential pair is V_{DD}-(V_{OD1}+$V_{SG2,3}$). Therefore it is not possible to use the same supply for DAC and comparator. In other words rail-to-rail conversion is not implementable here.

So, 1.2v and 0.7v is chosen as the supply of comparator and DAC respectively. Nodes VD4 and VD5 are precharged "low" while V_{out} and $\overline{V_{out}}$ precharged "high". In a rising clock edge M1 reaches triode region, therefore, between M10 and M11, the one that has higher gate charge state, switch on earlier and the corresponding drain will be low.

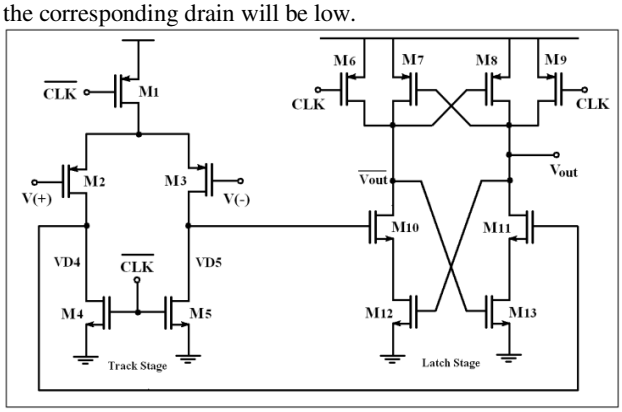


Fig. 3: Track-and-Latch comparator inspired from [9]

IV. Post Layout Simulation Results and Disscussion

The contribution of this paper is extracting post layout simulations, which distinguishes with previous paper [10]. The delay of the ADC from rising edge of comparator clock until appearance of valid data at the output of the encoder is almost 28ns that means the speed of this ADC could be almost 36MS/s. Indeed for lower frequency the power consumption will be lower; this fact is illustrated in Fig. 4. It shows that this ADC consumes almost 18µW at 36MS/s and 110nW at 0.1MS/s. TABLE I summarizes the ADC performance. A histogram (code density) test with a periodic ramp input has been done to calculate INL and DNL with 0.1LSB DNL resolution [11]. The simulated DNL and INL in the ADC are given in Fig. 5. They show that the DNL and INL of the proposed flash ADC are 0.64LSB and 0.62LSB, respectively. Such a small DNL guarantees that no missing code will occur. These values increased comparing to [10] mainly due to parasitic effects of the layout. Dynamic Characteristics of the ADC are processed with MATLAB performing Fast Fourier Transform (FFT). A 0.6MHz input sinusoidal wave is sampled with 10MS/s sampling frequency. Results for SINAD, THD, ENOB and SFDR are summarized in TABLE I.

For low power ADC applications in biomedical implants and wireless sensor network, power consumption is a critical characteristic. One ordinary used figure of merit (FOM) for power dissipation is the energy consumed per conversion step. Proposed figure of merit by [12] FOM=P/ (2^N.f_S), is more common, where P is the power consumption, N is the number of bits and f_S is the sampling frequency. Based on this FOM, the designed ADC consumes only 31fJ per conversion step. This value is rather higher than the FOM presented in [10] mainly due to parasitic effects of the layout.

TABLE I: Performance of the ADC

Resolution	4 bit
Maximum Sampling rate	36MS/s
Input Voltage Range	0-0.7v
DNL	0.64LSB
INL	0.62LSB
SFDR	44.4dB
SINAD	35.5dB
THD	-35.7dB
ENOB	3.8
Supply Voltage Resistor String Comparator Digital Blocks	 0.7 V 1.2 V 0.5 V
power consumption @ Maximum Speed	 18µW
FOM @ Maximum Speed	31 fJ/conversion
Technology	0.18µm CMOS

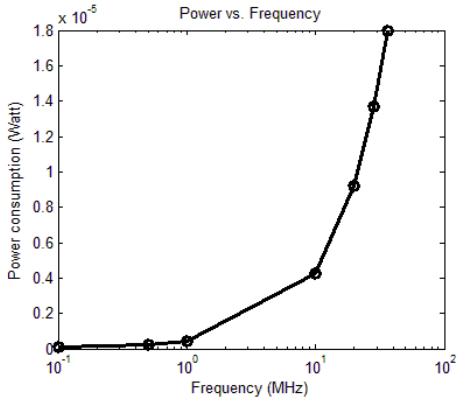


Fig. 4: Power consumption vs. frequency

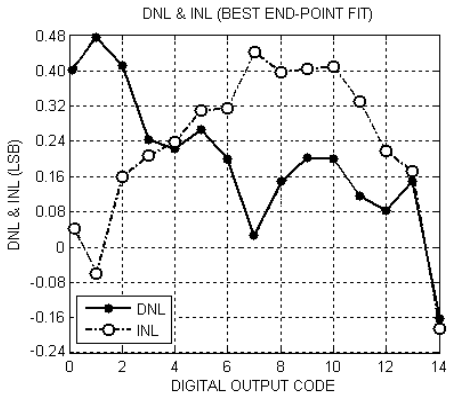


Fig. 5: Differential nonlinearity and Integral nonlinearity @ 10MS/s

Fig. 6 illustrates the variation of the FOM versus sampling frequency. It shows for lower and higher the FOM increases. So, there exists an optimum frequency with a given FOM. This optimum frequency is around 20MHz. TABLE II makes a comparison with some other recently published ADCs. Among these papers, only [13] is a low power SAR ADC and the others are low power flash ADCs. Fig.7 shows the layout of the ADC including resistor-string, digital blocks, decoupling capacitors and pads.

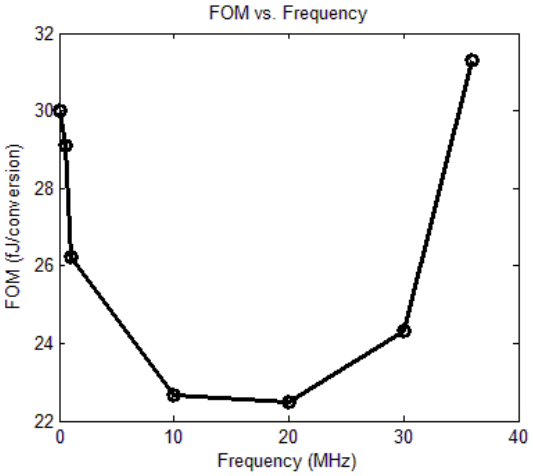


Fig. 6: FOM vs. sampling frequency

V. Conclusion

In this paper, an ultra-low power Flash ADC with a dual-stage was proposed and designed with 4 bit resolution. The maximum achievable speed was 31MS/s, at the cost of only 18µW power dissipation. To reduce the total power consumption this flash ADC uses 16 high value resistors as resistor string (ladder) and a comparator with no static current were used. The encoder and other digital blocks enjoy low supply voltage to further reduce the power consumption. The results show that the proposed ADC has higher speed with almost the same power consumption in comparison with SAR ADC family. Also, this architecture occupies less area than conventional SAR architectures such as fully charge redistribution.

TABLE II Comparison table

	[13]	[7]	[14]	[15]	[16]	This work
Power Cons.	13.4 µW	115 µW	0.82 mW	19.2 mW	1.93 mW	18 µW
Bit	10	4	5	6	4	4
Sampling Freq.	137kS/s	50 MS/s	1 GS/s	800 MS/s	1.2 GS/s	36 MS/s
Tech (µm)	0.18	0.13	0.09	0.35	0.09	0.18
DNL (LSB)	0.56	2.75	0.3	0.2/ -0.7	0.57	0.64
INL (LSB)	0.38	2	0.3	0.4/ -0.8	0.39	0.62
FOM fJ/conv	95	143	39	375	100	31

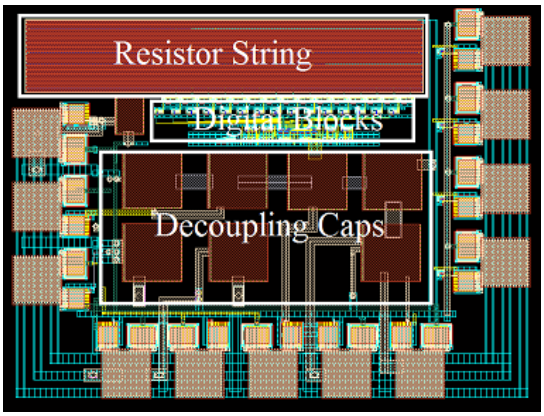


Fig. 7: Layout of the Flash ADC.

Acknowledgment

The authors would like to express their sincere thanks to Professor M. B. Nejad for his unsparing supports.

References

[1] Finkenzeller, *RFID Handbook*, 2nd ed., John Wiley & Sons Ltd, 2003.
[2] D. J. Yeager, et al, "Wirelessly-Charged UHF Tags for Sensor Data Collection," *IEEE Conference on RFID, Las Vegas, USA,* Apr. 2008.
[3] H. Deng, at al "Design of Sensor-Embedded Radio Frequency Identification (SE-RFID) Systems" *IEEE International Conference on Mechatronics and Automation* Jun. 2006.
[4] J. Marjonen, R. Alaoja, H. Ronkainen, M. Aberg, "Low power successive approximation A/D converter for passive RFID tag sensors," *IEEE* Apr. 2006.
[5] A. P. Sample, et al, "Design of an RFID-Based Battery-Free Programmable Sensing Platform," *IEEE Trans. On Instrumentation and Measurement,* Nov 2008, Vol. 57, No. 11.
[6] David A. Johns, Ken Martin. *Analog Integrated Circuit Design,* 1998.
[7] Mohammad Masoumi, Erik Markert, Ulrich Heinkel, Georges Gielen, "Ultra Low Power Flash ADC for UWB Transceiver Applications" *IEEE* 2009.
[8] M. Baghaei-Nejad, D. S. Mendoza, Z. Zou, S. Radiom, G. Gielen, Z. Li-Rong, and H. Tenhunen. "A remote-powered RFID tag with 10Mb/s UWB uplink and -18.5dBm sensitivity UHF downlink in 0.18um CMOS," *IEEE International Solid-State Circuits Conference - Digest of Technical Papers*, ISSCC 2009, pp. 198-199,199a.
[9] M. V. Elzakker, et al, "A 1.9µW 4.4fJ/Conversion-step 10b 1MS/s Charge-Redistribution ADC," *IEEE International Solid-State Circuits Conference*, 2008.
[10] I. Kianpour, M. Baghaei-Nejad, S. M. A. Zeinolabedin, Z. Li-Rong, "A subthreshold ultra low power 22fJ/conversion flash ADC for RFID sensing applications", IEEE, Electrical Engineering (ICEE), 19th Iranian Conference on Electrical Engineering (ICEE), 2011.
[11] W. Kester, "Data Conversion Handbook", Analog Devices, Inc., 2005. ISBN: 0-7506-7841-0.
[12] B. Murmann, M. Steyaert, A.H.M. Roermund, J.H. van Huijsing, "Limits on ADC Power Limitation," *Springer, Analog Circuit Design* 2006.
[13] K. Hoonki, M. YoungJae, K. Yonghwan, and K. Soowon, "A Low Power Consumption 10-bit Rail-to-Rail SAR ADC Using a C-2C Capacitor Array", *IEEE*, 2008.
[14] B. Verbruggen, et al, "A 2.2 mW 1.75 GS/s 5 Bit Folding Flash ADC in 90 nm Digital CMOS", *IEEE JOURNAL OF SOLID-STATE CIRCUITS*, March 2009, Vols. VOL. 44, NO. 3.
[15] Wen-Ta Lee, Po-Hsiang Huang, Yi-Zhen Liao and Yuh-Shyan Hwang, "A New Low Power Flash ADC Using Multiple-Selection Method", *IEEE*, 2007.
[16] L. Ying-Zu, L. Yu-Chang and C. Soon-Jyh, "A 0.35-1 V 0.2-3 GS/s 4-bit Low-Power Flash ADC for a Solar-Powered Wireless Module", *IEEE*, 2010.

SESSION 5

DISTRIBUTED SYSTEMS AND SOFTWARE ENGINEERING

An Overview of the IEEE 802.15.4e Standard

Erico Meneses Leão
Doctoral Programme in Informatics Engineering
Faculty of Engineering, University of Porto
Porto, Portugal
Email: pro12001@fe.up.pt

***Abstract*—The use of wireless sensor networks (WSN) in real-time applications is an attractive research topic due to its wide application range such as environment monitoring, biomedical, military and industrial applications. In last years, the IEEE 802.15.4 is a standard designed for Low-Rate Wireless Personal Area Networks (LR-WPANs). Up to this moment, one of the main research topics has been energy efficient operation, whereas real-time aspects were not a primary concern. Thus, the IEEE 802.15.4e Task Group (TG4e) was created to define a MAC amendment to the existing standard 802.15.4-2011. The intent of this amendment is to enhance and add functionality to the 802.15.4-2011 MAC to better support the industrial applications (critical requirements i.e. low latency, robustness and determinism). This paper presents an overview of the IEEE 802.15.4e standard.**

***Index Terms*—Wireless Sensor Network, industrial application, real-time, IEEE 802.15.4 standard, IEEE 802.15.4e standard.**

I. Introduction

Automation activities are essential for the competitiveness increase in all industrial sectors. Industrial automation can be characterized as a set of techniques that enable the construction of active subsystems with capability to interact with the industrial processes for control, monitoring, and supervision purposes [1].

According to [2], an increasing number of industrial applications are focusing on wireless networks as a core technology. Wireless technologies have also been identified as a very attractive option for industrial and factory automation, distributed control systems, automotive systems and other kinds of networked embedded systems [3].

Within this context, wireless sensor systems can revolutionize industrial processing and help industry meet the demands of increased competitiveness. Wireless sensor technology offers reliable, autonomous process control to improve product quality, increase yield and reduce costs. Thus, the use of Wireless Sensor Networks (WSN) in the industry is an active topic of research.

One of the more frequently used wireless Sensor Networks technologies is the IEEE 802.15.4. The IEEE 802.15.4-2011 [4] is a standard designed for Low-Rate Wireless Personal Area Networks (LR-WPANs), which focus on short-range operation, low-data rate, energy-efficiency, and low-cost implementations.

The IEEE 802.15.4 standard has become a recognized industry standard and provides a specification for both the Physical Layer (PHY) and Medium Access Control (MAC) sublayers [2]. One of the main design goals of this standard has been energy efficient operation, whereas hard real-time aspects were not a primary concern [2], [5].

Thus, the IEEE 802.15.4e Task Group (TG4e) [6] was created to define an amendment to enhance the IEEE 802.15.4-2011 standard. This amendment, namely IEEE 802.15.4e standard [7], introduces additional medium access control (MAC) mechanism and frame formats that allow devices to support a wide range of industrial and commercial applications.

This paper presents the main features of the IEEE 802.15.4e standard, with a special emphasis on the enhancements of the MAC access mechanisms.

The remainder of this paper is organized as follows: Section II summarize the basics in IEEE 802.15.4 standard. Section III summarize the main characteristics and enhancements of IEEE 802.15.4e standard. Section IV reviews some relevant studies about IEEE 802.15.4 and IEEE 802.15.4e standards. Finally, the paper is concluded in section V.

II. Overview of the IEEE 802.15.4 Standard

The IEEE 802.15.4-2011 standard [4] defines the physical layer (PHY) and medium access control (MAC) sublayer specifications for low-data-rate wireless connectivity. This section reviews the main characteristics of the IEEE 802.15.4 standard, including physical details, CSMA-CA, contention based or contention free access mechanism.

A. General Characteristics

A Low-Rate Wireless Personal Area Network (LR-WPAN) is a simple, low-cost communication network that allows wireless connectivity in applications with limited power and relaxed throughput requirements. Some of the capabilities provided by the 802.15.4 standard are as follows [4]:

- Star or peer-to-peer operation;
- Unique 64-bit extended address or allocated 16-bit short address;
- Optional allocation of guaranteed time slots (GTSs);
- Carrier sense multiple access with collision avoidance (CSMA-CA) or ALOHA channel access;
- Fully acknowledged protocol for transfer reliability;
- Low power consumption;
- Energy detection (ED);
- Link quality indication (LQI).

B. Physical Layer

The physical (PHY) layer is responsible for various tasks, as follows: activation and deactivation of the radio transceiver, energy detection (ED), link quality indicator (LQI), clear channel assessment (CCA) for carrier sense multiple access with collision avoidance (CSMA-CA), channel frequency selection, data transmission and reception and precision ranging for ultra-wide band (UWB) PHYs [4].

The devices can operate in one or several bands using one of the modulation and spreading formats available in the standard.

C. Network Topologies

The IEEE 802.15.4 standard defines two types of devices: a full-function device (FDD) and a reduced-function device (RFD). An FFD is a device that is capable of serving as a personal area network (PAN) coordinator, while an RFD is a device that is not capable of serving as PAN coordinator. A Wireless Personal Area Network (WPAN) is compose by multiple FFD and RFD devices, with a FFD designated as PAN coordinator.

The standard can operate in either of two topologies: the star topology or the peer-to-peer topology. Figure 1 shows the IEEE 802.15.4 standard topologies.

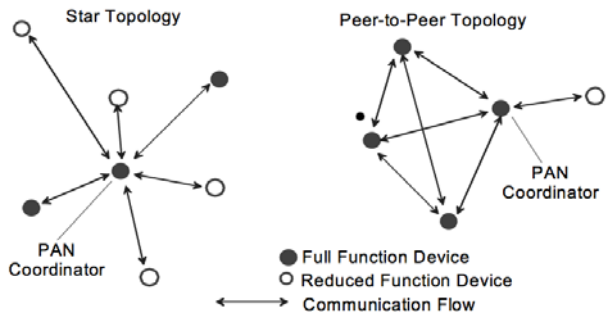


Fig. 1. Star and Peer-to-Peer Topologies [4]

In the star topology, all communication is realized between devices and the PAN coordinator (only FFD device). The PAN coordinator is the primary controller and it is used to initiate, terminate, or route communication around the network. According to the standard, all devices operating on a network of either topology have unique addresses (extended addresses). A device will use either the extended address for direct communication within the PAN or the short address that was allocated by the PAN coordinator when the device was associated.

In the peer-to-peer topology, any device may communicate with any other device as long as they are in the range of one another. This topology allows more complex network formations to be implemented, such as mesh networking topology. It may also allow multiple hops to route messages from any device to any other device on the network. Such functions can be added at the higher layer, but are not part of this standard [4].

D. Medium Access Control (MAC) Layer

The Medium Access Control (MAC) layer (sublayer) provides an interface between the physical layer and upper layers of the WPAN. The MAC sublayer addresses all access to physical radio channel and it is responsible for the following tasks: beacon synchronization and management, channel access, association and disassociation support, guaranteed time slot (GTS) management, frame validation, and acknowledged frame delivery.

The carrier sense multiple access - collision avoidance (CSMA-CA) is the adopted channel access method. The MAC protocol of IEEE 802.15.4 standard supports two operation modes (channel access mechanism): nonbeacon-enable and beacon-enable.

1) Nonbeacon-enabled mode: The nonbeacon-enabled mode use an unslotted CSMA-CA channel access mechanism. This mechanism inserts a random waiting time (*backoff periods*) between 0 and $2^{BE} - 1$, where BE *Backoff Exponential* shall be initialized to the value of *macMinBE*. According to the IEEE 802.15.4 standard, if the channel is found to be idle, following the random backoff, the device transmits its data. If the channel is found to be busy following the random backoff, the device waits for another random period before trying to access the channel again.

2) Beacon-enabled mode: The beacon-enabled mode uses a slotted CSMA-CA channel access mechanism. This mode uses beacon frames transmitted by the coordinator. The beacon messages are used to delimit a well-defined structure called superframe. Figure 2 shows the superframe structure.

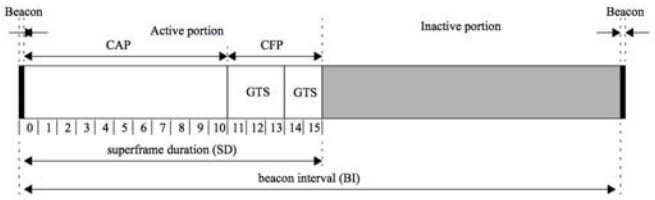


Fig. 2. Superframe structure [4]

The superframe can have an active portion and an inactive portion. In the inactive portion all nodes (including the coordinator) can sleep. The active portion of each superframe is divided into 16 time slots and is composed of three parts: a beacon, a contention access period (CAP) and a contention-free period (CFP).

In the CAP period, if a device wishes to communicate, it will have to contend with other devices using slotted CSMA-CA mechanism to access the channel. For low latency applications or applications that require real-time operation, so the CFP period is introduced. In the CFP period, the coordinator allocates the *guaranteed time slots* (GTS) for the devices. In these slots, the nodes can transmit data without contending for the channel access. The PAN coordinator can allocate up to seven GTSs, and a GTS is allowed to occupy more than one slot period [4].

The superframe structure is described by the values of *macBeaconOrder* (BO) and *macSuperframeOrder* (SO). They define the *beacon interval* (BI) and *superframe duration* (SD), respectively. The BI determines the interval at which the coordinator shall transmit its beacon frames. The SD determines the length of the active portion of the superframe. The BI and SD are defined as follows:

$$BI = aBaseSuperframeDuration.2^{BO}(symbols) \quad (1)$$

$$SD = aBaseSuperframeDuration.2^{SO}(symbols) \quad (2)$$

where $0 \leq SO \leq BO \leq 14$.

3) CSMA-CA Algorithm: The CSMA-CA algorithm is based on three variables: *Number of Backoff* (NB), *Contention Window* (CW), and *Backoff Exponent* (BE). According to IEEE 802.15.4 standard, NB (initial value is 0) is the number of times the CSMA-CA algorithm was required to backoff while attempting the current transmission. CW is the contention window length, defining the number of backoff periods that need to be clear of channel activity before starting the transmission. BE is the backoff exponent, which is related to how many backoff periods a device shall wait before attempting to assess the channel.

Firstly, the CSMA-CA algorithm checks the *battery life extension* field to define the *macMinBE* attribute. If the *battery life extension* is true, then the value given to *macMinBE* is minimal(2, *macMinBE*); otherwise, the value is *macMinBE*. After, the algorithm waits for a random counter (backoff period) between 0 and $2^{BE}-1$. When this backoff period expires, the algorithm executes a *Clear channel assessment* (CCA). If the channel is found busy, the NB and BE parameters are incremented (BE shall not exceed *aMaxBE*). If the number of backoff exceeds the *macMaxCSMABackoffs* of channel, the transmission fails.

Once the channel is idle, CW is decreased. The CCA operation is repeated until $CW = 0$. If the channel is still idle, the transmission is dispatched. According to the standard, if the number of backoff periods is greater than the remaining number of backoff periods in the CAP, the MAC sublayer shall pause the backoff countdown at the end of the CAP and resume it at the start of the CAP in the next superframe. Figure 3 illustrates the CSMA-CA algorithm.

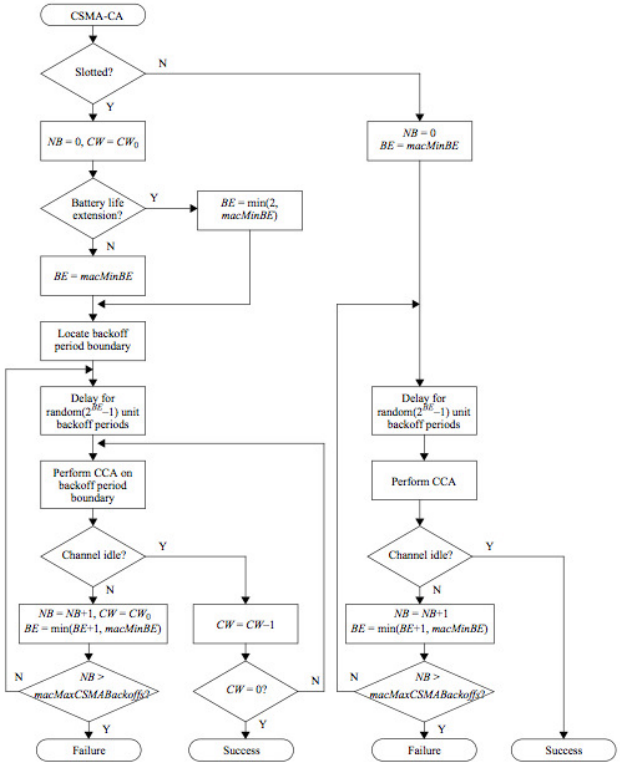


Fig. 3. CSMA-CA Algorithm [4]

III. Overview of the IEEE 802.15.4e Standard

The IEEE 802.15.4e Task Group (TG4e) [6] [8] was created to define a MAC amendment to the existing standard 802.15.4-2011 [4]. The intent of this amendment was to enhance and add functionality to the 802.15.4-2011 MAC, in order to better support the industrial market requirements. Industrial applications have critical requirements (i.e. low latency and robustness) and determinism that are not adequately addressed by IEEE Std 802.15.4-2011 [7]. According to the standard, there are two categories of MAC enhancements, as follows: behaviors to support application such as process and factory automation, and general functional improvements.

In order to achieve hardware compatibility, the IEEE 802.15.4 PHY layer was totally preserved. Thus, IEEE 802.15.4e implements only a MAC protocol modification (it does not require any change to the hardware).

The IEEE 802.15.4e standard defines some MAC behavior modes, as follows:

- Deterministic and synchronous multi-channel extension (DSME): for general industrial and commercial application domains;
- Timeslotted channel hopping (TSCH): for application domains such as process automation;
- Low Latency deterministic networks (LLDN): for application domains such as factory automation;
- Radio frequency identification blink (RFID): for application domains such as item and people identification, location, and tracking.

In this paper, we introduce the DSME and TSCH modes.

A. Deterministic and synchronous multi-channel extension (DSME)

The Deterministic and synchronous multi-channel extension (DSME) mode provides the following features [7]:

- Multi-channel, multi-superframe, mesh extension to GTS for deterministic latency, flexibility, and scalability;
- Group acknowledgment option for higher reliability and efficiency;

- Distributed beacon scheduling and distributed slot allocation for robustness and scalability;
- Two channel diversity modes (channel adaptation and channel hopping) for robustness and higher reliability even in dynamic channel conditions.

The IEEE 802.15.4 standard provides only up to seven *guaranteed time slots* (GTS) to support applications requiring deterministic delay. Moreover, GTSs are restricted to use a single channel [4], [7]. The DSME mode enhances IEEE Std 802.15.4-2011 in two important directions: extension of number of GTS timeslots and the number of frequency channels used [7]. To accommodate these enhancements, the DSME mode adopts a versatile multi-superframe structure (extension for the IEEE 802.15.4 superframe).

1) Multi-superframe structure: The multi-superframe is defined by the coordinator. The multi-superframe comprises multiple superframes. The beginning is dedicated for beacon transmission and the reimander of superframe consists of a Contention Access Period (CAP) and a Contention-Free Period (CFP). Figure 4 shows the DSME multi-superframe structure.

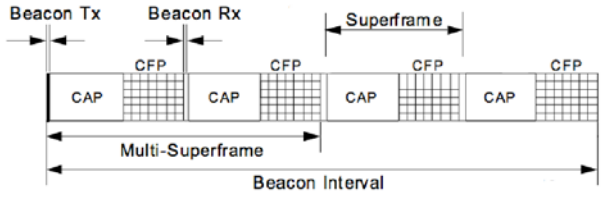


Fig. 4. DSME Multi-Superframe Structure

The multi-superframe structure is described by the values of *macBeaconOrder* (BO) and *macSuperframeOrder* (SO), and *macMultiSuperframeOrder* (MO). The values of the *beacon interval* (BI), *superframe duration* (SD), and *multi-superframe duration* (MD) are related as follow:

$$BI = aBaseSuperframeDuration.2^{BO}(symbols) \quad (3)$$

$$SD = aBaseSuperframeDuration.2^{SO}(symbols) \quad (4)$$

$$MD = aBaseSuperframeDuration.2^{MO}(symbols) \quad (5)$$

where $0 \leq SO \leq MO \leq BO \leq 14$. Thus, the number of superframes in a multi-superframe is defined by 2^{MO-SO} and the number of multi-superframes in a beacon interval is 2^{BO-MO}. An example, with BO=6, SO=3, and MO=5, there are four superframes in a multi-superframes and two multi-superframe in the Beacon Interval.

2) Channel diversity: Wireless communications are vulnerable to mutual channel interferences and channel fading [9]. The DSME protocol provides two channel diversity methods: channel adaptation and channel hopping. In channel adaptation, a pair of devices (source and destination) can allocate DSME-GTSs in a single channel or in different channels. In other words, the devices can switch to other frequency channels, with better link quality. In channel hopping mode, a pair of devices can hop over predefined frequency channel list (regardless of channel conditions). According to [9], considering that adjacent links can use different channels by employing channel diversity, aggregate network throughput increases significantly in multi-hop environments by maximizing the number of simultaneous transmission in the PAN.

Figure 5 shows an example of channel usage DSME-GTS in channel hopping. In this example, each DSME-GTS is represented by a tuple (timeslot, channel). Thus, for device A, the DSME-GTSs are (1,6), (2,5), (3,4), (4,3), (5,2), (6,1) and (7,6). For device B are (1,5), (2,4), (3,3), (4,2), (5,1), (6,6) and (7,5).

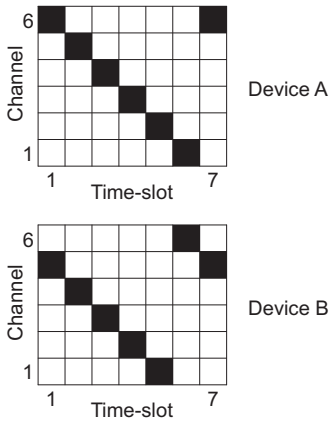


Fig. 5. Example of channel usage of DSME-GTS

According to [7], the transmitting device shall switch to the channel used by the receiving device in order to send a data frame. If the receiving device receives the data frame successfully, it sends an Acknowledgement (ACK) frame to the transmitting device on the same channel.

B. Timeslotted channel hopping (TSCH)

The Timeslotted channel hopping (TSCH) mode uses time synchronized communication and channel hopping to provide network robustness through spectral and temporal redundancy. The TSCH can be used to form any topology from a star to a full mesh [7].

In the TSCH mode, the superframe is replaced with a slotframe. A slotframe is a group of timeslots repeating in time. The number of timeslots determines how often each timeslot repeats, thus setting a communication schedule for nodes that use the timeslots. A pair of devices can exchange a frame and, optionally, an acknowledgment in each timeslot.

Figure 6 illustrates a slotframe. In this case, the slotframe is composed by three timeslots. Thus, nodes A and B can communicate in the timeslot 0 and nodes B and C can communicate in the timeslot 2, for example.

It is possible the use of multiple slotframes to define a different communication schedule for various nodes. In this case, a device may participate in one or more slotframes simultaneously, and not all devices need to participate in

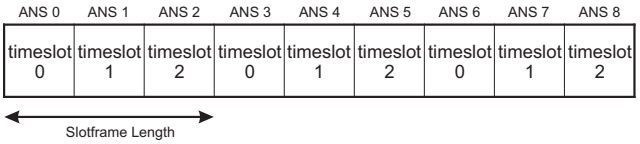


Fig. 6. Example of a slotframe

all slotframes [7]. Figure 7 shows an example of multiple slotframes.

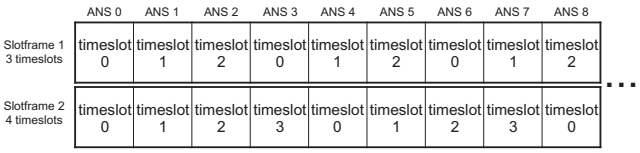


Fig. 7. Multiple slotframe

IV. Related Works

In paper [10], the authors take a brief overview of the PHY and MAC layer of the five standards (IEEE 802.15.4-2003, IEEE 802.15.4b-2006, IEEE 802.15.4a-2007, IEEE 802.15.4c-2009 and IEEE 802.15.4d-2009) and discuss what aspects led to these changes. Similarly, the authors of paper [11] give a brief overview of WPAN and wireless sensor network protocols, addressing the IEEE 802.15.4 standard.

The paper [8] proposes a *Multi-Factor Dynamic GTS Allocation Scheme* (MFDGAS) to improve the utilization of GTS bandwidth. According to authors, in the MFDGAS, the PAN coordinator determines the allocation of GTSs by taking the data size, delay time and the utilization of GTS time slot into consideration. The paper results show that the proposed GTS allocation scheme improves the performance of system throughput.

The literature proposes a few papers on the new IEEE 802.15.4e standard, because it is a recently published standard (April 2012). Some papers have been published based on the draft document in the new standard. In paper [2], the authors discuss the applicability of IEEE 802.15.4 for application in industrial automation and the necessity of a standard for LR-WPAN with real-time aspects. The paper shows the performance IEEE 802.15.4-based protocol variant in a comprehensive set of simulation experiments (industrial scenario).

In paper [9], the authors evaluate performance of time synchronized multi-channel (DSME) MAC protocol for wireless sensor networks and the results are compared with that of IEEE 802.15.4 slotted CSMA-CA MAC protocol in a beacon-enable PAN. The DSME is a MAC behavior mode defined in the IEEE 802.15.4e standard. The performance metrics used by authors were throughput and energy consumption.

The authors of paper [12] propose the adaptive slotted channel hopping (A-TSCH). The A-TSCH is an enhanced version of the TSCH aided by blacklisting technique. Blacklisting is a practice that excludes undesirable channels from the hopping sequence. According to authors, the A-TSCH protocol can significantly improve the reliability of channel hopping scheme and thus provide better protection from interference for wireless sensor networks.

Finally, in paper [13], the authors present an overview of IEEE 802.15.4e TSCH MAC protocol. Moreover, the authors conceived a novel traffic aware scheduling algorithm by exploiting graph theoretical arguments to support emerging industrial applications requiring low latency at low duty cycle and power consumption.

V. Conclusion

Currently, wireless sensor networks have drawn an important attention from research groups due to its application range. This context, the industrial applications are adopting the wireless sensor network technology.

One of the most frequently used wireless sensor network technologies is IEEE 802.15.4. The IEEE 802.15.4-2011 is a standard designed for Low-Rate Wireless Personal Area Networks (LR-WPANs), which focus on short-range operation, low-data rate, energy-efficiency, and low-cost implementations. However, industrial applications have critical requirements such as low latency, robustness and determinism. The IEEE 802.15.4 standard does not addresses these requirements appropriately.

This paper presents a brief overview of the newly presented IEEE 802.15.4e standard. The standard was created to define an amendment to the popular IEEE 802.15.4-2011 standard. This amendment introduces additional medium access control (MAC) behaviors and frame formats that allow devices to support a wide range of industrial and commercial applications.

Thus, we studied the main MAC protocols of the standard, DSME and TSCH, and identified their main characteristics and applications. We observed that characteristics of the DSME and TSCH protocols (i.e. extension of number of GTS timeslots, several frequency channels, channel adaptation and channel hopping) eliminate IEEE 802.15.4 limitations.

The compatibility of IEEE 802.15.4e with the IEEE 802.15.4 standard and a wide application variety are fundamental characteristics to the success of the new standard.

References

[1] E. Leao, L. Guedes, and F. Vasques, "An event-triggered smart sensor network architecture," in *Industrial Informatics, 2007 5th IEEE International Conference on*, vol. 1, june 2007, pp. 523 –528.

[2] F. Chen, R. German, and F. Dressler, "Towards IEEE 802.15.4e: A study of performance aspects," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on*, 29 2010-april 2 2010, pp. 68 –73.

[3] A. Willig, "Recent and emerging topics in wireless industrial communications: A selection," *Industrial Informatics, IEEE Transactions on*, vol. 4, no. 2, pp. 102 –124, may 2008.

[4] "IEEE standard for local and metropolitan area networks–part 15.4: Low-rate wireless personal area networks (LR-WPANs)," *IEEE Std 802.15.4-2011 (Revision of IEEE Std 802.15.4-2006)*, pp. 1 –314, 5 2011.

[5] F. Chen, R. German, and F. Dressler, "Qos-oriented integrated network planning for industrial wireless sensor networks," in *Sensor, Mesh and Ad Hoc Communications and Networks Workshops, 2009. SECON Workshops '09. 6th Annual IEEE Communications Society Conference on*, june 2009, pp. 1 –3.

[6] (2012) IEEE 802.15.4 WPAN task group tg4e website. [Online]. Available: http://www.ieee802.org/15/pub/TG4e.html

[7] "IEEE standard for local and metropolitan area networks–part 15.4: Low-rate wireless personal area networks (LR-WPANs) amendment 1: MAC sublayer," *IEEE Std 802.15.4e-2012 (Amendment to IEEE Std 802.15.4-2011)*, pp. 1 –225, 16 2012.

[8] C.-L. Ho, C.-H. Lin, W.-S. Hwang, and S.-M. Chung, "Dynamic GTS allocation scheme in IEEE 802.15.4 by multi-factor," in *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2012 Eighth International Conference on*, july 2012, pp. 457 –460.

[9] W.-C. Jeong and J. Lee, "Performance evaluation of IEEE 802.15.4e dsme mac protocol for wireless sensor networks," in *Enabling Technologies for Smartphone and Internet of Things (ETSIoT), 2012 First IEEE Workshop on*, june 2012, pp. 7 –12.

[10] N. Salman, I. Rasool, and A. Kemp, "Overview of the IEEE 802.15.4 standards family for low rate wireless personal area networks," in *Wireless Communication Systems (ISWCS), 2010 7th International Symposium on*, sept. 2010, pp. 701 –705.

[11] S. Wang, Y. Zhang, Z. Liu, W. Zhou, and D. Liu, "A brief study on low-rate wireless personal network," in *Systems and Informatics (ICSAI), 2012 International Conference on*, may 2012, pp. 1360 –1363.

[12] P. Du and G. Roussos, "Adaptive time slotted channel hopping for wireless sensor networks," in *Computer Science and Electronic Engineering Conference (CEEC), 2012 4th*, sept. 2012, pp. 29 –34.

[13] M. R. Palattella, N. Accettura, M. Dohler, L. A. Grieco, and G. Boggia, "Traffic aware scheduling algorithm for reliable low-power multi-hop IEEE 802.15.4e networks," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2012 IEEE 23rd International Symposium on*, sept. 2012, pp. 327 –332.

A Fault Localization Approach to Improve Software Comprehension

Alexandre Perez, Rui Abreu
Department of Informatics Engineering
Faculty of Engineering, University of Porto
Porto, Portugal
alexandre.perez@fe.up.pt, rui@computer.org

***Abstract*—Program understanding is one of the most important, but also amongst the most time consuming phases of software maintenance tasks. Currently, there are tools to assist in program understanding by means of dynamic analysis, but suffer from scalability issues when used in large, real-world software applications. We propose an approach, coined PANGOLIN, that makes use of concepts and techniques from the software fault localization field, which were proven to be effective even when debugging large applications in resource-constrained environments. PANGOLIN analyses the program at hand by exploiting run-time information from test case executions and computes the similarity of each component to the functionality being maintained, helping software engineers in understating how a program is structured and what the functionality's dependencies are. A case study with the open-source application Rhino is presented, demonstrating the efficacy of PANGOLIN in locating the components that should be inspected when changing a certain functionality.**

Keywords*—*Software Engineering, Software Evolution and Maintenance, Fault Diagnosis, Program Comprehension.

I. INTRODUCTION

Software maintenance is a crucial part of software engineering. The need to add or change new features to existing software applications is becoming more and more prevalent. Furthermore, the ever increasing complexity of software systems and applications renders software maintenance even more challenging.

One of the most daunting tasks of software maintenance is to understand the application at hand, in order to evolve it [1]. During this program understanding task, software engineers try to find a way to make both the source code and the overall program functionality more intelligible. One of these ways is to create a mental map of the system structure, its functionality, and the relationships and dependencies between software components [2], [3].

To fully understand how a software application behaves, software engineers need to thoroughly study the source code, documentation and any other available artifacts. Only then does the engineer gain sufficient understanding of the application at hand, enabling him/her to conduct the desired maintenance or evolution tasks. This *program comprehension* (also known as *program understanding*) phase is thus very resource and time consuming. In fact, studies show that up to 50% of the time needed to complete maintenance tasks is spent on understanding the software application and gaining sufficient knowledge to change the desired functionality [1]. Therefore, if tools are able to aid software engineers in understanding program functionality, considerable gains in maintenance efficiency can be achieved [1].

Currently, there are several tools and techniques that use dynamic analysis approaches to provide visualizations of the software system, identifying their components and their relationships [4], [5], [6]. However, these approaches, due to their dynamic nature and having to deal with vast amounts of data, have often scalability issues [7].

To address some of the issues of past approaches, we propose a new dynamic method, coined PANGOLIN, that exploits some of the concepts and techniques used in the software fault localization field. Fault localization techniques exploit run-time information (*program spectra*) of test cases to calculate the likelihood of each component being faulty, and were shown to be efficient, even for large, resource-constrained environments [8]. Our PANGOLIN approach leverages these concepts to provide an efficient dependency analysis for software comprehension that does not have the same scalability hinderances as other related tools.

The paper makes the following contributions:

- We propose PANGOLIN, an approach that, similarly to fault localization techniques, exploits run-time information from system executions to identify dependencies between components, helping software engineers in understanding how a program is structured.
- We provide a toolset integrated with the fault localization tool GZoltar [9], that uses PANGOLIN to provide a visualization of correlated and dissociated components of an application functionality.
- A case study with a large, real-world, software project, demonstrating the efficacy of our approach in locating the components that should be inspected when evolving/changing a certain functionality.

To the best of our knowledge, an approach that leverages coverage-based fault localization techniques to improve program understanding has not been described before.

The remainder of this paper is organized as follows. In Section II we introduce some concepts relevant to this paper, namely program spectra and fault localization. Section III will present our PANGOLIN approach. In Section IV we discuss the results of the application of the PANGOLIN approach to the Rhino project. We provide an overview of the related work

and how it compares to PANGOLIN in Section V. Finally, in Section VI, we conclude and discuss future work.

II. PRELIMINARIES

In this section, program spectra will be introduced, and its use in fault localization will be detailed. After that, an approach to visualize diagnostic reports will be presented.

A. Program Spectra

A program spectrum is a characterization of a program's execution on an input collection [10]. This collection of data consists of counters of flags for each software component, and is gathered at runtime. Software components can be of several detail granularities, such as classes, methods or lines of code. A program spectrum provides a view on the dynamic behavior of the system under test [11].

Recording program spectrum is a very lightweight analysis method when compared to other dynamic run-time methods (such as, *e.g.*, dynamic slicing [12]). In order to obtain information about which components were covered in each execution, the program's source code needs to be instrumented, similarly to what happens in code coverage tools [13]. This instrumentation will monitor each component and register those that were executed.

Many types of program spectra exist. This paper focuses on registering whether a component is touched or not during a certain execution, so binary flags can be used for each component. This particular form of program spectra is also called hit spectra [11].

B. Fault Localization

Spectrum-based Fault Localization (SFL) is a statistical debugging technique that, for each software component, calculates the likelihood of it being faulty [14]. It exploits information from passed and failed system runs. A passed run is a program execution that is completed correctly, and a failed run is an execution where an error was detected [8]. The criteria for determining if a run has passed or failed can be from a variety of different sources, namely test case results and program assertions, among others. The execution information gathered for each run is their hit spectrum.

The hit spectra of N runs constitutes a binary $N \times M$ matrix A, where M corresponds to the instrumented components of the program. Information of passed and failed runs is gathered in an N-length vector e, called the error vector. The pair (A, e) serves as input for the SFL technique, as depicted in Fig. 1.

$$\begin{array}{c} \\ N \text{ spectra} \end{array} \begin{array}{c} \overset{M \text{ components}}{\overbrace{\phantom{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \end{bmatrix}}}} \\ \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NM} \end{bmatrix} \end{array} \quad \begin{array}{c} \text{error} \\ \text{vector} \\ \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} \end{array}$$

Fig. 1. Input to SFL.

With this input, the next step consists in identifying what columns of the matrix A (*i.e.*, components) resemble the error vector the most. This is done by quantifying the resemblance between these two vectors by means of *similarity coefficients* [15].

Several similarity coefficients do exist [8]. Examples of similarity coefficients are shown below, namely the Jaccard coefficient s_J used in the Pinpoint tool [16], and the s_T coefficient used in the Tarantula[1] tool [14], [17]:

$$s_J(j) = \frac{n_{11}(j)}{n_{11}(j) + n_{01}(j) + n_{10}(j)} \tag{1}$$

$$s_T(j) = \frac{\frac{n_{11}(j)}{n_{11}(j)+n_{01}(j)}}{\frac{n_{11}(j)}{n_{11}(j)+n_{01}(j)} + \frac{n_{10}(j)}{n_{10}(j)+n_{00}(j)}} \tag{2}$$

where $n_{pq}(j)$ is the number of runs in which the component j has been touched during execution ($p = 1$) or not touched during execution ($p = 0$), and where the runs failed ($q = 1$) or passed ($q = 0$). For instance, $n_{11}(j)$ counts the number of times component j has been involved in failed executions, whereas $n_{10}(j)$ counts the number of times component j has been involved in passed executions. Formally, $n_{pq}(j)$ is defined as:

$$n_{00}(j) = |\{i \mid a_{ij} = 0 \wedge e_i = 0\}| \tag{3}$$

$$n_{01}(j) = |\{i \mid a_{ij} = 0 \wedge e_i = 1\}| \tag{4}$$

$$n_{10}(j) = |\{i \mid a_{ij} = 1 \wedge e_i = 0\}| \tag{5}$$

$$n_{11}(j) = |\{i \mid a_{ij} = 1 \wedge e_i = 1\}| \tag{6}$$

One of the best performing similarity coefficients for fault localization is the Ochiai coefficient [18]. The fault localization tools Zoltar [19] and GZoltar[2] [9] use the Ochiai coefficient to quantify the resemblance to the error vector. This coefficient was initially used in the molecular biology domain [20], and is defined as follows:

$$s_O(j) = \frac{n_{11}(j)}{\sqrt{(n_{11}(j) + n_{01}(j)) \cdot (n_{11}(j) + n_{10}(j))}} \tag{7}$$

The calculated similarity coefficients rank the software components according to their likelihood of containing the fault. This is done under the assumption that a component with a high similarity to the error vector has a higher probability of being the cause of the observed failure. A list of the software components, sorted by their similarity coefficient, is then presented to the developer. This list is also called *diagnostic report*, and helps developers prioritize their inspection of software components to pinpoint the root cause of the observed failure.

[1] Available at http://pleuma.cc.gatech.edu/aristotle/Tools/tarantula/

[2] Available at http://wwww.gzoltar.com

	mid() {	Runs						
	int x,y,z,m;	1	2	3	4	5	6	s_O
1:	read("Enter 3 numbers:",x,y,z);	●	●	●	●	●	●	0.41
2:	m = z;	●	●	●	●	●	●	0.41
3:	if (y<z) {	●	●	●	●	●	●	0.41
4:	if (x<y)	●	●			●	●	0.50
5:	m = y;		●					0.0
6:	else if (x<z)	●				●	●	0.58
7:	m = y; //BUG	●				●		0.71
8:	} else {			●	●			0.0
9:	if (x>y)			●	●			0.0
10:	m = y;			●				0.0
11:	else if (x>z)				●			0.0
12:	m = x;							0.0
13:	}							0.0
14:	print("Middle number is:",m);	●	●	●	●	●	●	0.41
	} Error vector:	✓	✓	✓	✓	✗	✓	

Fig. 2. Example of SFL technique with Ochiai coefficient (adapted from [17]).

In Fig. 2 it is shown an example of the SFL technique, using the Ochiai coefficient. To improve this example's legibility, the coverage matrix and the error detection vector were transposed. The detail granularity for each component of the hit spectra in the example is the line of code. In this example, the system under test is a function named `mid()` that reads three integer numbers and prints the median value. This program contains a fault on line 7 - it should read `m = x;`.

Six test cases were run, and their coverage information for each line of code can be seen to the right. At the bottom there is also the pass/fail status for each run - which corresponds to the error detection vector e. According to this pass/fail status, the test 5 fails, and the other ones pass. Then, the similarity coefficient was calculated for each line using the Ochiai coefficient (Equation 7). These results represent the likelihood of a certain line containing a fault. The bigger the coefficient, the more likely it is of a line containing a fault. Therefore, these coefficients can be ranked to form an ordered list of the probable faulty locations.

In this specific example, the highest coefficient is in line 7 - the faulty line. The SFL technique has successfully performed the fault localization. Although small, this example serves well to demonstrate how SFL works.

C. Diagnostic Report Visualization/Inspection

In order to improve the intuitiveness of the diagnostic report generated by SFL tools, interactive visualization techniques have been proposed, such as the sunburst visualization, available in the GZoltar tool [9], which is a fault localization plugin for the Eclipse[3] integrated development environment (IDE). Besides fault localization, GZoltar also offers tools for test suite minimization and prioritization, but those are beyond the scope of this paper.

In the sunburst visualization, each ring denotes a hierarchical level of the source code organization. From the inner to the outer circle, this visualization presents projects, packages, files, classes, methods and lines of code.

Navigation in this visualization is done by clicking on a component, which will display all the inner components of the selected one. As an example, if a user clicks in a class, all of the class methods will be displayed. Users may also zoom in/out and pan to analyze in detail a specific part of the system. Any inner component can also be set as the new root of the visualization, and only that component's sub-tree is displayed. This operation is called a *root change*.

The color of each component in the visualization represents its likelihood of being faulty. Ranging from bright green if the similarity is close to zero; to bright red if the component's similarity to the error vector is close to 1.

An example of the use of the sunburst visualization to debug a real application can be seen in Fig. 3. The figure shows the sunburst visualization of the open-source project NanoXML[4] after injecting a fault (so that some tests would fail), and running GZoltar's fault localization. As can be seen, the sunburst narrows down the faulty candidates (about 5 methods), lessening the effort required to locate the fault.

After fixing the fault and re-running GZoltar, the visualization becomes as shown in Fig. 4. As there is no fault (exercised by the test cases, *i.e.*, no test case fails), all software components are bright green.

The effectiveness of these fault localization techniques and hierarchical visualizations was also demonstrated in a user study [21], where 30 computer science students and researchers from 3 different Universities used GZoltar to debug a fault that was injected in the NanoXML project (the same fault presented above).

[3] Available at http://www.eclipse.org

[4] Available at http://devkix.com/nanoxml.php

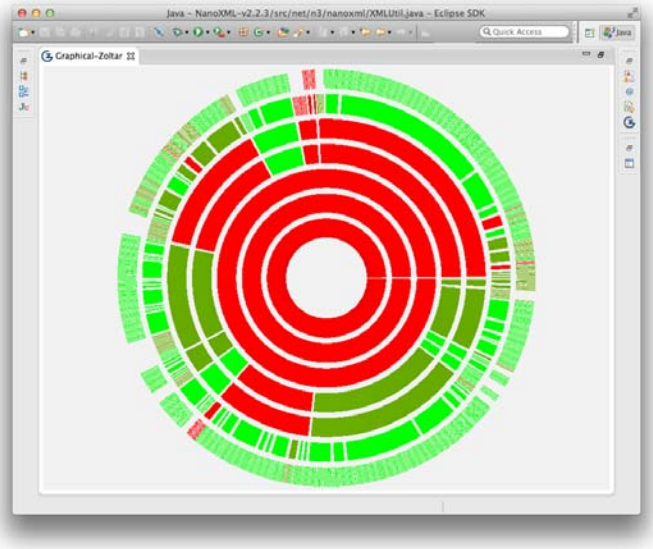


Fig. 3. Sunburst visualization of the NanoXML project with an injected fault.

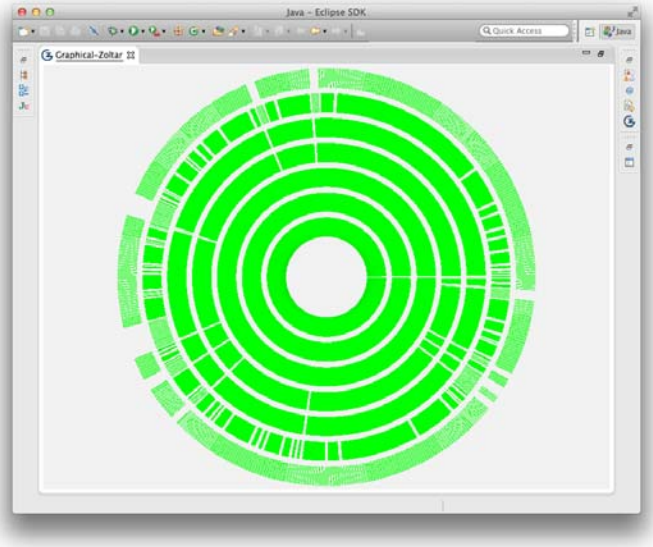


Fig. 4. Sunburst visualization of the NanoXML project with no observed failures.

This study shows that, without any knowledge of the system under test, 85% of the subjects were able to locate the fault under just 20 minutes. Furthermore, 70% of the subject were even able to fix the injected fault under the same amount of time.

None of the subjects from the group that did not use GZoltar was able to locate the fault under 20 minutes. In fact, in the post-experimental survey, the subjects stated that they would need much more time to debug the application.

III. Approach

In this section, a methodology to use coverage-based fault localization techniques in the fields of software evolution and maintenance, coined PANGOLIN, is proposed.

Maintenance and software evolution are becoming more and more prevalent. As companies change their business models and strategies, it is likely that their supporting systems and applications need to follow suit and change or add additional functionality.

The cost of developing new systems suited to answer the new demands is often prohibitive. A cheaper alternative is to change the existing applications. However, this alternative also poses significant challenges to developers, as most implementation details may have not been thoroughly documented. For instance, dependencies between components may be difficult to be identified [3].

Run-time analysis techniques similar to those used in software fault localization may help developers in understanding the inner workings of a particular software application and identifying dependencies between components that need to be changed.

A. Concepts & Definitions

In order to use these techniques some concepts and definitions need to be established. The first one is the notion of a feature.

Definition 1 *A **feature** is the source code portion that implements a certain functionality. It may contain one or more components.*

When trying to evolve/modify a certain feature f, we are interested in the relationships and interactions between f and other components in the source code. As such, one should not use the error vector e, input to traditional SFL, as seen in section II-B, to compute the similarity coefficient. Instead, an evolution vector ev_f should be used.

Definition 2 *The **evolution vector** ev_f is an N-length binary vector. In this vector, a given position i is true* (i.e., *set as* 1*), if the i^{th} test run has touched the feature f.*

As such, the input for this technique is now the pair (A, ev_f), and can be seen in Fig. 5.

$$\begin{array}{c} \\ N \text{ spectra} \end{array} \begin{array}{c} \overset{M \text{ components}}{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NM} \end{bmatrix}} \end{array} \quad \begin{array}{c} \text{evolution} \\ \text{vector} \\ \begin{bmatrix} ev_{f_1} \\ ev_{f_2} \\ \vdots \\ ev_{f_N} \end{bmatrix} \end{array}$$

Fig. 5. Input to the coverage-based approach.

Definition 3 *A component j is **correlated** with f if its similarity with f is close to* 1*. This means that when f is executed, j is likely to be executed as well.*

When evolving a feature f, it is important to inspect its correlated components because they may either call f or be

called by f, and thus may need to be modified in accordance with the changes made to f.

Definition 4 *A component j is **dissociated** to f if its similarity with f is close to 0. This means that when f is executed, j is not likely to be touched by the execution.*

In contrast to correlated components, if a component is dissociated to f, it does not need to be inspected when f is modified.

B. PANGOLIN *Algorithm*

The approach to compute the similarity of all components to the feature to be evolved, coined PANGOLIN, is detailed in Algorithm 1. In this section, this algorithm is explained in detail.

Algorithm 1 PANGOLIN algorithm.

Input: Program $\mathcal{P}$, set of test cases $\mathcal{T}$, and feature $\mathcal{F}$
Output: Report $\mathcal{R}$

```
1:  N ← |T|
2:  M ← NUMCOMPONENTS(P)
3:  R ← ∅
4:  A ← EXEC(P, T)
5:  ev ← UPDATE(A, F)
6:  for j = 0 → M do
7:      n00(j), n01(j), n10(j), n11(j) ← 0
8:  end for
9:  for i = 0 → N do
10:     for j = 0 → M do
11:         if A[i, j] = 0 ∧ ev[j] = 0 then
12:             n00(j) ← n00(j) + 1
13:         else if A[i, j] = 0 ∧ ev[j] = 1 then
14:             n01(j) ← n01(j) + 1
15:         else if A[i, j] = 1 ∧ ev[j] = 0 then
16:             n10(j) ← n10(j) + 1
17:         else if A[i, j] = 1 ∧ ev[j] = 1 then
18:             n11(j) ← n11(j) + 1
19:         end if
20:     end for
21: end for
22: for j = 0 → M do
23:     R[j] ← s_O(n00(j), n01(j), n10(j), n11(j))
24: end for
25: return R
```

The inputs for the PANGOLIN algorithm are as follows:

- The program under evaluation $\mathcal{P}$.
- A test suite $\mathcal{T}$.
- A feature to be evolved $\mathcal{F}$. It is a subset of $\mathcal{P}$'s components.

The output is the report $\mathcal{R}$, which is a list of components, each containing its similarity coefficient to the feature under consideration.

First, the variables N and M are created, corresponding to the size of the test suite and the number of components that exist in the program $\mathcal{P}$, respectively (lines 1 and 2).

After that, the test suite $\mathcal{T}$ of the program $\mathcal{P}$ is executed (line 4). The program spectra matrix that resulted from this execution is stored in matrix A. This matrix contains the execution traces for every test in $\mathcal{T}$. Next, the evolution vector is calculated, by comparing each line of the matrix A to the feature vector $\mathcal{F}$.

Afterwards, the for loops in lines 6 and 9 are used to fill the occurrence function n_{pq} (as defined in section II-B) for each component of program $\mathcal{P}$.

Finally, for each component, the similarity is calculated with the Ochiai coefficient (see Equation 7), and stored in the report $\mathcal{R}$, which is returned afterwards. It is worth noting that, unlike what happens in SFL, the report $\mathcal{R}$ does not have to be sorted. This is because the report will not be inspected as a ranking by the user. The report will be used solely to populate the sunburst visualization.

C. Complexity Analysis

This section presents the space and time complexity analysis for the PANGOLIN approach detailed in the sections above.

As for the space complexity, the generated program spectra matrix A has a complexity of $O(M \cdot N)$. The evolution vector and the report $\mathcal{R}$ complexities are $O(N)$ and $O(M)$, respectively. Therefore, the worst case space complexity is $O(M \cdot N + N + M)$.

The time complexity is as follows. Assuming that all the test cases in suite $\mathcal{T}$ are executed and take the same amount of time to execute, the complexity of this test execution step is $O(N)$. As for the evolution vector computation, its worst case time complexity is $O(M \cdot N)$. The computation of the n_{pq} occurrence function also has a complexity of $O(M \cdot N)$. Finally, the Ochiai coefficient calculation to populate the report $\mathcal{R}$ is $O(M)$. The worst case time complexity is $O(M \cdot N + N + M)$.

IV. CASE STUDY

In this section, the PANGOLIN technique and the sunburst visualization applied to the evolution of a real-world application will be presented.

The GZoltar toolset was used as the base framework for the case study. The PANGOLIN algorithm presented in section III was implemented in this tool, so that it uses evolution vectors to calculate what components are correlated and dissociated to a feature under evolution.

A. Rhino

The software application under consideration for this case study is the open-source project Rhino[5]. Rhino is a Javascript engine written entirely in Java, and is managed by the Mozilla Foundation[6]. It comprises 28 packages, 433 classes and 75170 lines of code. Furthermore, this project contains 448 unit tests, written using the JUnit[7] framework.

[5]Available at https://developer.mozilla.org/en-US/docs/Rhino
[6]http://www.mozilla.org/foundation/
[7]Available at https://github.com/KentBeck/junit

In this case study, the output of the modified GZoltar tool will be presented, for two features being considered for evolution: one responsible for validating regular expressions and other responsible for the creation of execution information (*i.e.*, context generation). For each feature, the results obtained will be described and discussed.

B. Regular Expression Validation

The first feature under consideration for evolution in this case study is a method that validates regular expressions. For instance, if the regular expression syntax used within Rhino needed to be changed, this validation method would need some modifications, as well as eventual components that call or are called by the method.

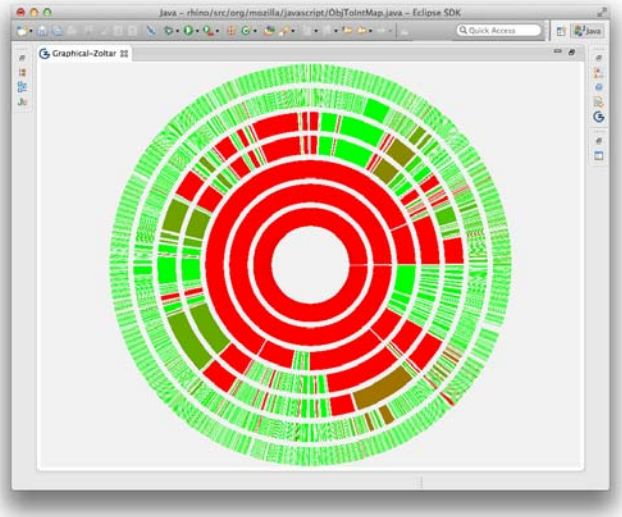


Fig. 6. Overall visualization while evolving the regular expression validation feature.

Fig. 6 depicts the GZoltar overall sunburst visualization, fully expanded. As mentioned on section II-C the similarity is color coded, ranging from bright green for components with a similarity coefficient of 0 (*i.e.*, dissociated components), to bright red for similarity coefficients of 1 (*i.e.*, correlated).

The visualization shown in the figure presents the user with an overall sense of what components or regions of the code one must inspect so that the changes made to the regular expression validation method do not break any functionality of the application. This visualization also shows code regions that do not need any sort of inspection (in bright green), because the functionality is dissociated with the feature considered for evolution. There is no communication between the validation method and those areas of the source code.

In order to provide a more detailed view of specific parts of the system, the *root change* functionality can be employed. As stated in section II-C, the *root change* operation allows for a inner component to be set as the new root of the visualization. An example of the root change operation can be seen in Fig. 7. This root change was done to a class that was correlated with the feature under evolution. In one of the class methods, there are two correlated lines of code. These two lines need to be

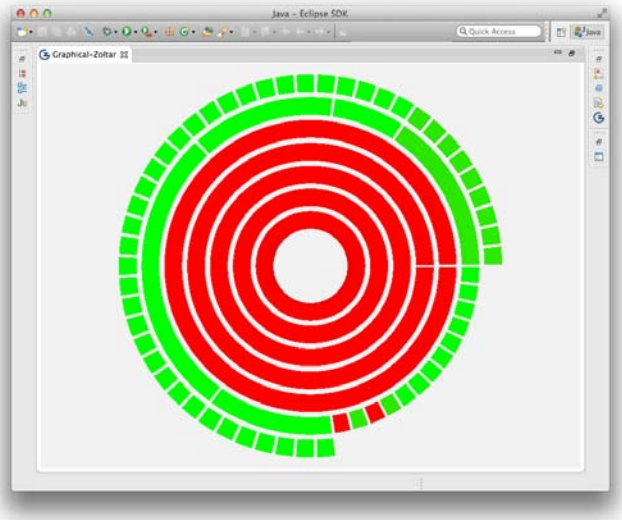


Fig. 7. Root change visualization while evolving the regular expression validation feature.

Fig. 8. Root change visualization while evolving the regular expression validation feature.

inspected, because they can call or be called by the feature. However, the user can change them safely without worrying about breaking any functionality, because these lines are only executed when the feature under consideration is also executed. This happens because, in correlated components, n_{10} equals 0. The remaining components, be it either other lines of code from the same function, or the other methods of the class that is zoomed in, do not need to be inspected by the user. This is because the components are dissociated to the feature under evolution.

Other example of a root change operation can be seen in Fig. 8. In this example, one of the methods of the expanded class has a darker green color associated to it. In fact, its similarity coefficient of roughly 0.45. This means that this

method is not only used in executions involving the feature under consideration. It can be, for example, an utility method. In these cases, changes in these methods can lead to unwanted failures of other functionality. So thorough testing is likely necessary to ensure that no functionality was broken.

C. Context Generation

The second feature under consideration for evolution in this case study is the context generation method. This method is responsible for creating a *context* structure, that will store information about the executing Javascript code. One example of the information stored in a context is the script's call stack.

Fig. 9. Overall visualization while evolving the context generation feature.

Fig. 10. Root change visualization while evolving the context generation feature.

In Fig. 9 it is shown the overall visualization for the context generation method dependencies. Overall, the context generation has less correlated components when compared to the example in the previous section.

Fig. 10 depicts a root change operation on one of the classes of the source code. One of the methods of that class hasa brownish red color associated to it. Its similarity coefficient is around 0.95. This means that the method is mainly used when a context is created. However, in executions where contexts are not created, this method is still called, although very rarely. This should be an indicator that the method is being incorrectly called, or that its functionality is not properly modularized. This is also the case for components with low similarity scores (*e.g.*, such as 0.05).

V. Related Work

Various techniques and tools were developed as a result of several years of research into trace visualization. This section provides an overview of the related work in this area.

De Pauw *et al.* [22], [4] have developed a tool - coined Jinsight - for visually exploring a program's runtime behavior. Although this tool was shown to be useful for program comprehension, scalability concerns render the tool impractical for use in large applications.

Reiss [5] states that execution traces are typically too large to be visualized and understood by the user. As such, Reiss proposed a way to select and compact trace data to improve the visualization's intelligibility. Live run-time visualizations have also been proposed (rather than postmortem trace visualizations) as a way to reduce overheads [23], but making it harder to visualize entire executions.

Greevy *et al.* [6] proposes a 3D visualization of the run-time traces of a software system. Greevy displays the amount of information about a component as a tower whose height is influenced by the amount of instances created. The main objective of this technique is to determine which system regions are involved in the execution of a certain feature, but the visualization is not trivial to grasp.

Cornelissen *et al.* [24], [25] developed a tool - coined Extravis - that visualizes execution traces by employing two synchronized views: a circular bundle view for structural elements and an interactive overview via a sequence view. Its effectiveness was also demonstrated for three reverse engineering contexts: exploratory program comprehension, feature detection and feature comprehension.

The PANGOLIN approach proposed in this paper differs from the related work because of the low overhead necessary to compute the similarity coefficients for all the application's components. Another advantage is that it can not only pinpoint what components should be inspected and what components can be completely disregarded when evolving a feature, but also can warn about the existence of functionality that is not properly modularized.

VI. Conclusions & Future Work

Maintenance and software evolution are becoming more and more important to companies, as their business models

and strategies change. This change will likely influence their supporting systems and applications. Therefore, changes and additions to a software program's functionality are inevitable. As most implementation details are often poorly documented, component dependencies may be a crucial hinderance in the evolution of software, especially for older (heavily patched) and/or legacy systems.

This paper proposes an automated run-time method to lessen the effort required by developers in identifying dependencies between features that they are trying to evolve (or maintain) and the rest of the program. This approach, coined PANGOLIN, is based on statistics-based methods used in software fault localization. These methods exploit run-time information (*program spectra*) of test cases to calculate the likelihood of each component being faulty. In the field of software evolution, the resemblance to the code being evolved is calculated for all components. This can identify correlated components, that need to be changed because they may call or be called by the evolving component, and the dissociated components, that can be disregarded by the developer as they do not touch the functionality under consideration.

A case study with the open software project Rhino was performed. It demonstrated the efficacy of the PANGOLIN approach in pinpointing the components that should be inspected when evolving/changing a certain feature and those that can safely be disregarded. Furthermore, our approach was also able to warn about the possible existence of functionality that was not properly modularized.

As for future work, a user study will be conducted, asserting the effectiveness of the method described in this paper. Other visualization methods will also be considered.

REFERENCES

[1] T. A. Corbi, "Program understanding: challenge for the 1990's," *IBM Syst. J.*, vol. 28, no. 2, pp. 294–306, Jun. 1989.

[2] D. Lange and Y. Nakamura, "Object-oriented program tracing and visualization," *Computer*, vol. 30, no. 5, pp. 63 –70, May 1997.

[3] M. Renieris and S. P. Reiss, "Almost: exploring program traces," in *Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation in conjunction with the eighth ACM internation conference on Information and knowledge management*, ser. NPIVM '99. New York, NY, USA: ACM, 1999, pp. 70–77.

[4] W. De Pauw, D. Lorenz, J. Vlissides, and M. Wegman, "Execution patterns in object-oriented visualization," in *Proceedings Conference on Object-Oriented Technologies and Systems (COOTS98)*, 1998, pp. 219–234.

[5] S. P. Reiss and M. Renieris, "Encoding program executions," in *Proceedings of the 23rd International Conference on Software Engineering*, ser. ICSE '01. Washington, DC, USA: IEEE Computer Society, 2001, pp. 221–230.

[6] O. Greevy, M. Lanza, and C. Wysseier, "Visualizing live software systems in 3d," in *Proceedings of the 2006 ACM symposium on Software visualization*, ser. SoftVis '06. New York, NY, USA: ACM, 2006, pp. 47–56.

[7] A. Zaidman, "Scalability solutions for program comprehension through dynamic analysis," in *Software Maintenance and Reengineering, 2006. CSMR 2006. Proceedings of the 10th European Conference on*, Mar. 2006, pp. 4 pp. –330.

[8] R. Abreu, P. Zoeteweij, R. Golsteijn, and A. J. C. Van Gemund, "A practical evaluation of spectrum-based fault localization," *Journal of Systems and Software*, vol. 82, no. 11, pp. 1780–1792, 2009.

[9] J. Campos, A. Riboira, A. Perez, and R. Abreu, "GZoltar: an eclipse plug-in for testing and debugging," in *Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering*, ser. ASE 2012. New York, NY, USA: ACM, 2012, pp. 378–381.

[10] T. Reps, T. Ball, M. Das, and J. Larus, "The use of program profiling for software maintenance with applications to the year 2000 problem," in *Proceedings of the 6th European Software Engineering conference held jointly with the 5th ACM SIGSOFT international symposium on Foundations of software engineering*, ser. ESEC '97/FSE-5. New York, NY, USA: Springer-Verlag New York, Inc., 1997, pp. 432–449.

[11] M. J. Harrold, G. Rothermel, K. Sayre, R. Wu, and L. Yi, "An empirical investigation of the relationship between fault-revealing test behavior and differences in program spectra," *STVR Journal of Software Testing, Verification, and Reliability*, no. 3, pp. 171–194, Sep. 2000.

[12] B. Korel and J. Laski, "Dynamic program slicing," *Information Processing Letters*, vol. 29, no. 3, pp. 155–163, 1988.

[13] Q. Yang, J. J. Li, and D. Weiss, "A survey of coverage based testing tools," in *Proceedings of the 2006 international workshop on Automation of software test*, ser. AST '06. New York, NY, USA: ACM, 2006, pp. 99–103.

[14] R. Abreu, P. Zoeteweij, and A. J. C. van Gemund, "On the accuracy of spectrum-based fault localization," in *Proceedings of the Testing: Academic and Industrial Conference Practice and Research Techniques - MUTATION*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 89–98.

[15] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[16] M. Y. Chen, E. Kiciman, E. Fratkin, A. Fox, and E. Brewer, "Pinpoint: Problem determination in large, dynamic internet services," in *Proceedings of the 2002 International Conference on Dependable Systems and Networks*, ser. DSN '02. Washington, DC, USA: IEEE Computer Society, 2002, pp. 595–604.

[17] J. A. Jones and M. J. Harrold, "Empirical evaluation of the tarantula automatic fault-localization technique," in *Proceedings of the 20th IEEE/ACM international Conference on Automated software engineering*, ser. ASE '05. New York, NY, USA: ACM, 2005, pp. 273–282.

[18] R. Abreu, P. Zoeteweij, and A. J. C. v. Gemund, "An evaluation of similarity coefficients for software fault localization," in *Proceedings of the 12th Pacific Rim International Symposium on Dependable Computing*, ser. PRDC '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 39–46.

[19] T. Janssen, R. Abreu, and A. van Gemund, "Zoltar: A toolset for automatic fault localization," in *Automated Software Engineering, 2009. ASE '09. 24th IEEE/ACM International Conference on*, Nov. 2009, pp. 662 –664.

[20] A. da Silva Meyer, A. Garcia, A. de Souza, and C. de Souza, "Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (zea mays l.)," *Genetics and Molecular Biology*, vol. 27, pp. 83–91, 2004.

[21] J. Campos, *Regression Testing with GZoltar: Techniques for Test Suite Minimization, Selection, and Prioritization*. MSc Thesis, Faculdade de Engenharia da Universidade do Porto, 2012.

[22] W. De Pauw, R. Helm, D. Kimelman, and J. Vlissides, "Visualizing the behavior of object-oriented systems," in *Proceedings of the eighth annual conference on Object-oriented programming systems, languages, and applications*, ser. OOPSLA '93. New York, NY, USA: ACM, 1993, pp. 326–337.

[23] S. P. Reiss, "Visualizing java in action," in *Proceedings of the 2003 ACM symposium on Software visualization*, ser. SoftVis '03. New York, NY, USA: ACM, 2003, pp. 57–ff.

[24] B. Cornelissen, D. Holten, A. Zaidman, L. Moonen, J. J. van Wijk, and A. van Deursen, "Understanding execution traces using massive sequence and circular bundle views," in *Proceedings of the 15th IEEE International Conference on Program Comprehension*, ser. ICPC '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 49–58.

[25] B. Cornelissen, A. Zaidman, D. Holten, L. Moonen, A. van Deursen, and J. J. van Wijk, "Execution trace analysis through massive sequence and circular bundle views," *Journal of Systems and Software*, vol. 81, no. 12, pp. 2252 – 2268, 2008.

MatlabWeaver: an Aspect-Oriented approach for MATLAB

Tiago Carvalho, João Bispo, Pedro Pinto, João M. P. Cardoso
Informatics Engineering Department
Faculty of Engineering, University of Porto
Porto, Portugal
{tiagodrcarvalho, joaobispo}@gmail.com, ei07073@fe.up.pt, jmpc@acm.org

***Abstract*— MATLAB is a high-level language used by engineers to quickly develop and evaluate their solutions. The flexibility of MATLAB, however, comes at the cost of interpretation and lack of type/shape information. Moreover, the importance of trying different implementations requires maintenance of different application designs, which induces to a high percentage of tangled code. Aspect-Oriented Programming allows to increase the program modularity with cleaner code. MatlabWeaver is an Aspect-Oriented Programming approach, provided by MATISSE - a compiler framework for MATLAB code, which conveys information to the compiler regarding code modifications, types and array shape of the variables, allowing to try different code implementations, without changes in the original code.**

Keywords—Aspect-Oriented Programming, MATLAB, LARA, Compilers

I. INTRODUCTION

MATLAB[1] is a standard high-level language and an interactive numerical computing environment for many areas of application, including embedded computing. MATLAB is commonly used by engineers to quickly develop and evaluate their solutions. By including extensive domain-specific and visualization libraries, MATLAB substantially enhances engineers' productivity. Its flexibility, however, comes at the cost of interpretation and lack of type/shape information. In MATLAB the same identifier can be used to hold various data types and array shapes throughout the code execution. This makes it very handy for quick program prototyping but conspires against static analysis. The flexibility of *MATLAB* programs results in lower execution performance and the inability to directly use programs in production settings. This lack of performance is typically addressed by the development of an auxiliary reference implementation. This reference must then in turn be validated against the *MATLAB* code resulting on a lengthy and error prone process that further complicates the overall application development cycle and cost.

Aspect-Oriented Programming (AOP) is a programming paradigm whose objective is to increase modularity by separating cross-cutting concerns [1]. By separating secondary concerns from the core objective of the program using aspects results in cleaner code, easier concern analysis (the concerns are separated), easier for monitoring, tracing, debugging, etc. An AOP approach tries to aid programmers with mechanisms to achieve better modularity for their programs [1, 2].

AspectMATLAB[2] provides an aspect-oriented extension for MATLAB. Based on the aspect-oriented language AspectJ, it adds some distinctive features that are imperative for MATLAB programmers, such as: the ability to capture multidimensional array accesses and loops on MATLAB; and can recreate the action, using the join point information. However, AspectMATLAB does not provide selection over a specific point on the code, such as a specific statement or line, and lacks the definition of variable type and array shape[3, 4].

MATISSE [5], developed within the REFLECT project[3], provides a framework to help MATLAB development and to transform MATLAB code into equivalent C code. A core stage of MATISSE is the approach based on AOP concepts as a vehicle to convey information to the compiler regarding code modifications, such as code insertion for monitorization, debugging and instrumentation, and types and array shape of the variables. This allows rewriting applications for analysis, transformations, and code generation, trying different code implementations [5].

We present in this paper MatlabWeaver, the MATISSE core stage for representing AOP strategies. This approach consists on LARA, an AOP language designed for specifying non-functional requirements [6], and a developed weaver to specify those requirements. The system aids in the development of new systems by separating the secondary concerns, usually non-functional requirements, from the main application and allowing type and shape definition. Also, when converting code to a different target language, one can isolate the MATLAB code that is not intended to convert and stipulate specific code to use for the target language. This allows faster system development when trying different implementations and more flexibility for updates [5].

The rest of this paper is structured as follows: section II depicts MATISSE framework and its flow; in section III we describe MatlabWeaver, the weaving process in MATISSE using LARA; in section IV are exposed some case studies; section V presents the most related work; and we conclude the paper in section VI.

[1]Available at http://www.mathworks.com/products/matlab

[2]Available at http://www.sable.mcgill.ca/mclab/aspectmatlab/index.html

[3]Available at http://www.reflect-project.eu/

II. MATISSE

MATISSE is a compiler framework for generating low-level implementations from high-level specifications, guided by additional user constraints [5]. The current compiler aims at parsing high-level MATLAB code and generates C code for the low-level implementation. This framework backgrounds from the MATLAB prototype developed within the AMADEUS project[4], restructured to follow the REFLECT context. The new architecture allows the generation of C code from a large subset of MATLAB and a weaving stage to perform the aspect actions such as insertion of code, definitions of types and shapes, and specialization of the code based on default values [5].

Figure 1 illustrates the MATISSE framework. The first component, the Front-End, parses the MATLAB code and generates the intermediate representation (IR), which is used as input in the transformation tool, where one can apply strategies. MATISSE contains two Back-Ends that can generate MATLAB code or C code [5].

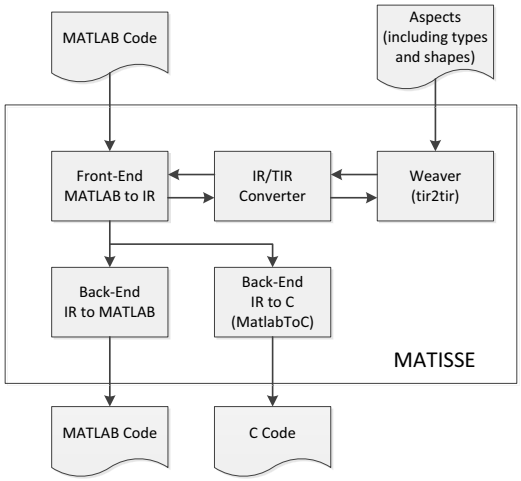


Figure 1. The MATISSE compiler framework.

Regarding the back-end, and when translating typeless/shapeless languages such as MATLAB to C, one of the important steps is the definition of specific types and in particular array variable shapes. For this, MATISSE contains a data-flow analysis approach as well as an external user-provided mechanism for static determination of types and shapes of variables [5].

MATISSE uses AOP concepts as a vehicle to convey information to the compiler regarding the types and array shape of the variables. The compiler uses the user-provided information and complements/checks its consistency against the information it can derive from its own analyses. More precisely, MATISSE uses the LARA AOP language [6-8] to provide that information [5].

III. MATLABWEAVER: A LARA-GUIDED WEAVING

The current implementation of MATISSE follows an aspect-oriented approach to perform source-to-source transformations in the high-level specification and for variable type and shape definitions for the low-level implementation. The stage responsible for carrying it out is the MatlabWeaver.

As mentioned before, our aspect-oriented approach consists on LARA, an AOP language designed for specifying non-functional requirements [6, 7], and a weaver that receives as input the intermediate representation of the aspects (Aspect-IR [8]) and a TOM [9] representation of the IR. LARA allows the developer to specify the non-functional requirements detached from the source code, which ones should influence the code and in what order. This allows one to try different designs very easy [7].

Aspect components rely on three important artifacts: the pointcut expression (resulting in a set of join points, each join point represent a location in the program, or more generically program execution points), the advice (where the action to perform to the join points in a pointcut is described), and the weaver (it applies the aspect modules over the code/representation of the application). Typically, the AOP has been extensively used for inserting monitoring code, but there have been many applications of the concept to parallel programming, to configuration, and to specialization.

Figure 2 illustrates the structure of a LARA aspect. Aspects can be decomposed in different aspectdefs for better modularity and aspects can invoke other aspects. Each aspectdef can have input and output parameters, static fields, and conditional checks to test if an aspect can run with the given inputs. For the main scope, LARA uses the JavaScript[5] syntactic and semantic elements, for the common code statements such as assignments, loops, conditional statements, among others; and weaving statements to select pointcuts and advise over the join points. The weaving statements comprise select, apply and condition statements. The select statement defines the join point selection over the source code.

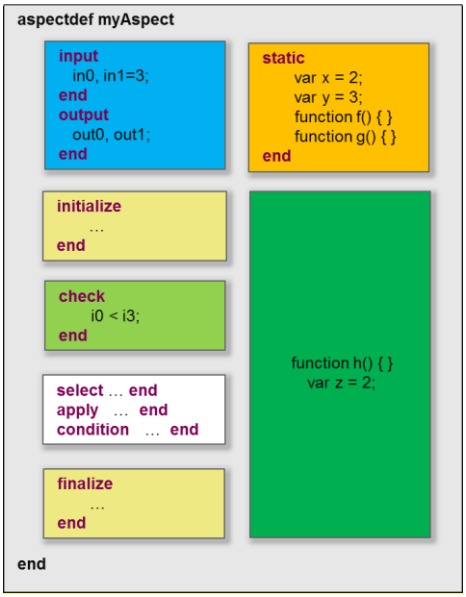


Figure 2. LARA aspect definition.

[4]Available at http://www.fe.up.pt/~specs/amadeus/

[5]Available at https://developer.mozilla.org/en/JavaScript

The apply statement can obtain information of each join point and advise over them. To filter the join points one can use the attributes filter in the select statement or use a condition statement over an apply.

LARA is an AOP language agnostic to the target language [6]. Therefore, its flexibility allows different input languages and weavers, provided that the join point model for the target language is specified. The aspects are defined the same way for any input language, which facilitates the aspect definition if one desires to apply a similar aspect to different target languages, such as MATLAB and C.

After writing the aspects, the LARA front-end (larac) parses the aspects and generates an Aspect-IR, according to the language specification [6, 8]. This Aspect-IR is the integration method used by LARA to link to the target weaver. The LARA interpreter tool (larai) can interpret the Aspect-IR, and can be used in different contexts. larai has been developed to simplify its integration in different weaving environments. It includes a java interface named IWeaver which requires the implementation of the basic functionalities, such as the select and action methods. The selection method uses no filter and requests all the join points the weaver can obtain. The filtering is subsequently dealt by the larai interpreter. This interface facilitates the development of weavers. It only needs the implementation of the required interface. The weaver integrated in MATISSE takes advantage of this functionality and implements the weaver interface for the TIR MATLAB representation.

The LARA language requires not only the aspect files but also a target language specification. This specification is both associated to the target language and to the weaver. Join point and attribute models specify the available points of interest in the code and their information, while the action model specifies the available code transformations.

A. *Join Point and Attribute Model*

The join point model defines the code structure in which the weaver works and the points of interest it can select and advice. The available model for MATLAB allows a vast set of points of interest in the code, including loops and uses and definitions of variables. Figure 3(a) depicts a parcel of the join point hierarchy which allows one to choose explicitly, for example, all the loops available in the source code, or to select implicitly one specific loop by its function parent or by its nested level.

This join point selection can be filtered by the information the weaver obtains from each join point. That information, usually called attributes, allows a fine grained selection over the join points. Some of the attributes in the MATLAB attribute model are illustrated in Figure 3(b).

Therefore, the model comprises the fundamental join points a MATLAB code may have. However, when the point of interest is outside the join point selection, one can use the statement join point, which generalizes any statement on the code and can be filtered, for instance, by the line number.Also, one can select a specific point or section in the code by defining specific annotations, named sections.

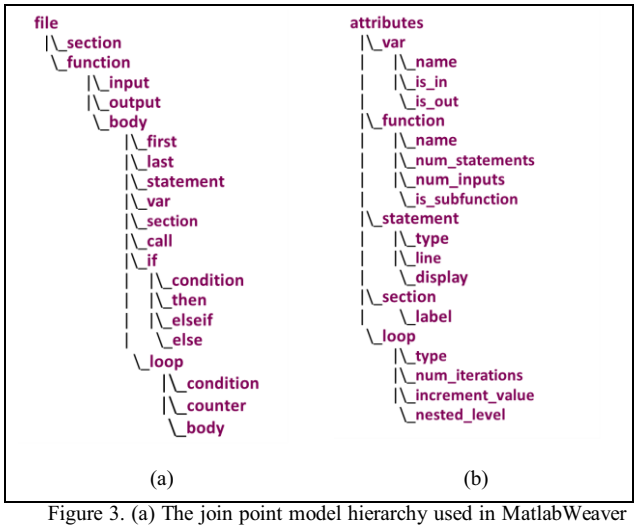


Figure 3. (a) The join point model hierarchy used in MatlabWeaver and (b) the attributes available for each join point.

Those sections can be selected the same way as the other type of join points and can be filter by the label given to the defined section. Figure 4 shows examples of the use of annotations in MATLAB and C code. The code contains specific code locations that could not be selected with the usual join points. A section join point using a single annotation refers to the code statement following the annotation.

```
function a = f1(b,c)
  a = 0;
  %@first
  b = b+1;
  a= f1(b,c);
  %@second
  a = f3(a,c);
  %@third
  a = a/2;
```

(a)

```
int f1(int b, int c) {
  int a = 0;
  #pragma label="first"
  b = b+1;
  a= f1(b,c);
  #pragma label="second"
  a = f3(a,c);
  #pragma label="third"
  a = a/2;
  return a;
}
```

(b)

Figure 4. Code example with annotations: (a) MATLAB code; (b) C code.

On the other hand, a section join point using two annotations (begin and end) refers to the code statements between those two annotations.

The following select statements represent the *selects* that captures those specific join points.

```
select section{label=="first"} end
select section{label=="second"} end
select section{begin=="first", end=="third"} end
```

Considering the MATLAB code in Figure 4 (a), the first select captures the MATLAB code statement:

```
b = b+1;
```

The second select captures the code statement:

```
a = f3(a,c);
```

And the last select captures the code section with the statements:

```
b = b+1;
a= f1(b,c);
#pragma label="second"
a = f3(a,c);
```

B. Action Model

The MATISSE weaver has currently two types of actions: code insertion and attribute definitions. Code insertion is a common action in the AOP languages. It is used for monitoring, profiling and tracing purposes. Since MATISSE can generate MATLAB and C code from a MATLAB source file, type and shape definitions are important aspects.

1) Code Insertion

The purpose of this action is to insert code before, after or instead of a statement or a code block, selected, according to the position parameter of the insert action. It allows writing non-functional requirements and better programming modularization by migrating the secondary functionalities to an aspect.

The following example depicts a MATLAB code with three sub-functions and a function call to one of the sub-functions. The objective is to replace the statement with the *dist* variable using a call to *euclidean* function by another function call and a display if the distance is less than a certain value.

```
...
  %@distance
  dist = euclidean(x,y);
...
function [ ecd ] = euclidean(x, y)
  ecd = sqrt(sum((abs(x-y)).^2));
end
function [ mhd ] = manhattan(x, y)
  mhd = sum((abs(x-y)));
end
```

The following aspect describes such secondary functionality, replacing the statement with a function call to the *manhattan* function.

```
  select section{label=="distance"} end
  apply
    insert around%{
      dist = manhattan(x,y);
      if dist < 100
         disp('manhattan distance less than 100!');}%;
  end
```

Applying this aspect to the source code changes it to a new source code containing the required changes. The following code depicts the code output by MATISSE.

```
// main MATLAB function
…
  dist = manhattan(x,y);
  if dist < 100
    disp('manhattan distance less than 100!');
  end
…
```

The code parameterization available in LARA allows the definition of aspects with parameters to be defined by the input arguments. With the aspect below it is possible to generate different codes using the same aspect with different values for the two input parameters. Some examples considering the *changeCalcDistance* aspect are depicted in Table I.

```
aspectdef changeCalcDistance
  input callFunc, threshold = 100 end
  select section{label=="distance"} end
  apply
    insert around%{
      dist = [[callFunc]](x,y);
      if dist < [[threshold]]
        disp('[[callFunc]] distance less than
[[threshold]]!');
      }%;
  end
end
```

TABLE I. EXAMPLES OF ASPECT CALLS WITH DIFFERENT INPUTS.

Aspect call	Code inserted
call changeCalcDistance("manhattan",150);	dist = manhattan(x,y); if dist < 150 disp('manhattan distance less than 150!');
call changeCalcDistance("euclidean",100);	dist = euclidean(x,y); if dist < 100 disp('euclidean distance less than 100!');

2) Define action

Define actions (identified in LARA as def) allows aspects to change some of the attributes. The available attributes which can be defined are the type and shape definition for variables and the default value for input variables.

The type definition is an important phase when transforming the MATLAB code into C, since MATLAB has no type and shape definition (default data-type is double) and the generated C code is more efficient when considering types and shapes known at compile time and in some cases the values may be represented with less costly data-types. For instance, the latnrm function, showed in Figure 5, when converted to C, all the types are considered with double precision.

```
function [outa]= latnrm(data, coefficient,
internal_state)
 NPOINTS = (64);
 ORDER = (32);
 bottom = (0);
 for i=1:1:NPOINTS
  top = data(i);
  ...
 end
end
```

Figure 5. Source code of *latnrm* function.

With LARA, the user can define the default type for all the variables and then specify a chosen type for a specific variable. The example in Figure 6 (a) depicts such aspect, considering the default type for all variables a fixed point with 32-bit word-length and 16 bits for the fraction, and an unsigned integer type of 16 bits for the NPOINTS and ORDER variables.

```
aspectdef VarTypeDefinition
  var uint16 = ["NPOINTS","ORDER"];
  allVars: select var end
  apply to allVars // default for all vars
    def type = " fixed(32,16)";
  end
  apply to allVars // variables with uint16
    def type = "uint16";
  end
  condition uint16.indexOf($var.name) > -1 end
end
```

(a)

```
scope latnrm {
   data: fixed(32,16)
   coefficient: fixed(32,16)
   internal_state: fixed(32,16)
   NPOINTS: uint16
   ORDER: uint16
   bottom: fixed(32,16)
   ...
}
```

(b)

Figure 6. (a) LARA aspect specifying types and (b) the resulting file with the types.

The MATISSE weaver generates a file with the types specified in LARA. This file is subsequently used by the code generators (e.g., ir2C and ir2matlab to generate, respectively, C code or MATLAB code). The code in Figure 6 (b) shows the file generated for the variables in *latnrm* function.

IV. Case Studies

As the general purpose of the MATISSE project is to translate code in a language to another target language more suitable for an embedded system implementation, the idea of deleting or adding code can help the translation process and can help the validation and exploration process. For example, functions used in MATLAB for displaying results are very useful in early development cycles, e.g., when exploring and validating solutions, but should not usually considered for the final embedded application.

We now show strategies regarding the use of MatisseWeaver in real applications, including specialization and code migration.

A. Variable Specialization: from constant to input

The latnrm function in Figure 5 contains two constant variables, NPOINTS and ORDER, which could become input parameters of the function, allowing the caller to this function to decide which values to use for these variables. Let's also consider that we intend to plot the outa matrix.

For this, the first step is to remove the assignments to these variables and then replace the function declaration to add the new input arguments. Then, to plot the outa matrix, we select the last statement of the outermost loop and insert, after that statement, the plot method.

These modifications can be achieved with the aspect depicted in Figure 7 (a). The woven MATLAB code generated by MATISSE according to the aspect in is illustrated in Figure 7 (b).

```
aspectdef LatnrmSpecialization
  select var end
  apply insert around%{}%; end
  condition($var.name=="NPOINTS"
      || $var.name=="ORDER") && $var.is_write end
  select function.declaration end
  apply
    insert around%{
      function [outa] = latnrm(data, coefficient,
        internal_state, NPOINTS, ORDER)}%;
  end
  select function.loop.last end
  apply
    insert after%{
      plot(1:size(outa, 2), outa);
      title('latnrm_out');}%;
  end
  condition $loop.is_outermost end
end
```

(a)

```
function [outa] = latnrm(data, coefficient,
internal_state, NPOINTS, ORDER)
   bottom = (0);
   for i=1:1:NPOINTS
      top = data(i);
      for j=2:1:ORDER
         left = (top);
         right = internal_state(j);
         internal_state(j) = bottom;
      ...
      outa(i) = sum;
      plot(1:size(outa, 2), outa);
      title( 'latnrm_out' );
   end
```

(b)

Figure 7. (a) Example of an aspect for the latnrm function and (b) Fraction of function latnrm after weaving (code highlighted corresponds to the woven code).

B. Variable Specialization: from input to constant

The purpose of the default value definition is to specialize code based on specific default values for certain parameters specified by the user. The MATLAB function parameters to be assigned to a default value are identified in the LARA aspect and defined with the default option. The strategy removes those parameters from the function declaration inputs and inserts the assignments (more precisely, the IR for those assignments) of these variables in the beginning of the IR representing the MATLAB function.

The following example shows the results of the strategy once it has been applied to a MATLAB function. The original MATLAB code contains the gridIterate which has six inputs. The last three inputs are intended to be removed from the function input parameters and placed in the beginning of the function.

```
function [v_old] = gridIterate(obstacle, v,
iter_max, nx, ny, nz)
 v0 = 0;
 v_end = -1;
 c = 1/6;
 for iter=1:iter_max
  ...
 end
 v_old = (v);
```

The corresponding LARA aspect responsible to specify this action is as follows. Based on the define action in LARA, the compiler removes nx, ny, and nz from the parameters of the function and inserts the 3 assignments nx = 32, ny = 64, and nz = 16.

```
aspectdef VarDefaultValue
  var defaults = {nx: 32, ny: 64, nz: 16};

  select function.input end
  apply def default = defaults[$input.name]; end
  condition $input.name in defaults end
end
```

Therefore, the generated MATLAB code includes only three inputs for the gridIterate function and default values for the selected former inputs.

```
function [v_old] = gridIterate(obstacle, v,
iter max)
    nx = (32);
    ny = (64);
    nz = (16);
    v0 = 0;
    v_end = -1;
    c = 1/6;
    ...
```

C. Code Migration

The purpose of this strategy is to migrate to aspects MATLAB code which is not intended to be converted into other programming language such as C. The strategy considers the secondary statements in a LARA aspect.

The following example shows how to use a LARA aspect to specify code modifications based on insert actions. For this example, the original MATLAB code for the harris function was used.

```
function [cim, …, csubp] = harris(im, …, disp)
  error(nargchk(2,5,nargin));
  if nargin == 4
    disp = 0;
  end
  if ~isa(im,'double')
    im = double(im);
  end
  subpixel = nargout == 5;

  % Compute derivatives
  [Ix, Iy] = derivative5(im, 'x', 'y');
  Ix2 = gaussfilt(Ix.^2,  sigma);
  Iy2 = gaussfilt(Iy.^2,  sigma);
  Ixy = gaussfilt(Ix.*Iy, sigma);
  % Compute the Harris corner measure
  % ….
  cim = (Ix2.*Iy2 - Ixy.^2)./(Ix2 + Iy2 + eps);

  if nargin > 2 % Perform nonmaximal suppression
    ...
  end
```

We migrate two sections of the code, which represent the secondary concerns related to error prevention and code not necessarily connected to the main concern, to an aspect, to the LARA aspect.

```
aspectdef MigrateHarris2ndConcerns
  select function.first end
  apply
    insert before %{
      error(nargchk(2, 5, nargin));
      ...
      subpixel = nargout==5;
    }%;
  end

  select function{label==”second”} end
  apply
    insert around %{
      if nargin > 2 % Perform nonmaximal suppression
        ...
      end
    }%;
  end
end
```

The resulting code presents the new MATLAB code of the harris function after migrating those blocks/statements to a LARA aspect. As can be seen, the source code becomes cleaner and easier to understand, while the secondary concerns are maintained in parallel on the aspect. Also important is the fact that the new original MATLAB code is the one that is intended to embedded implementation while the original MATLAB code was developed to be generic and as a model.

```
function [cim, …, csubp] = harris(im, …, disp)
    % Compute derivatives
    [Ix, Iy] = derivative5(im,  'x' ,  'y' );
    Ix2 = gaussfilt(Ix.^2, sigma);
    Iy2 = gaussfilt(Iy.^2, sigma);
    Ixy = gaussfilt(Ix.*Iy, sigma);
    % Compute the Harris corner measure
    % ….
    cim = ((Ix2.*Iy2-Ixy.^2)./(Ix2+Iy2+eps));
    %@second
    [r,c] = nonmaxsuppts(cim, radius, thresh);
```

The woven code obtained with the aspect and this new version of the code is the previous original code for the harris function. Using similar LARA aspects as the one used before in the context of the Euclidean example (see section III.B.1)) one can obtain different versions of the harris function to be implemented in an embedded system.

D. Aspects Metrics

To quantify the impact of MatlabWeaver over MATLAB applications, Table II measures some metrics retrieved in the analysis of the studied applications. The lines of code (LOC) are represented in three different ways: the first one represents the LOC of the original application; the second is the application after it is woven; and the third is the LOC of the aspect written in LARA. The number of switch points indicates a modification point in the code where aspect concerning code changes to functional code [10].

Aspect bloat refers to the aspect effectiveness with respect to the woven code. The metric is computed by the division of the concern LOC by the aspect LOC. Aspect bloat less than 1 means low aspect efficiency, since more code was written for the aspect than the concern; and a higher bloat means higher impact over the application and more reusability of the aspect.

Table II. Aspect metrics for the depicted strategies.

Plain code		**Weaved Code**						
Application Name	***Lines Of Code (LOC)***	***LARA Aspect Name***	***LOC (Woven Code)***	***LOC (Aspect)***	***Number of join points***	***Number of Switch Points***	***Aspect Bloat***	***Tangling Ratio***
latnrm	22	VarTypeDefinition	26	11	12	1	1,00	0,04
		LatnrmSpecialization	22	18	4	2	0,22	0,09
gridIterate	20	VarDefaultValue	23	6	3	1	0,67	0,04
harris	8[a]	MigrateHarris2ndConcerns	32	28	2	2	0,89	0,06

[a] LOC After Migrating the secondary concerns to the MigrateHarris2ndConcerns

The number of lines in the woven code is usually higher than the original code, since the aspects are inserting code with new secondary concerns. For the user, it is easier to focus the MATLAB code on the main concern and implement the secondary concerns in an aspect. As so, *harris* application is a moral example of how the secondary concerns pollute the code, since the original code, without the secondary concerns, is four times smaller than with the *MigrateHarris2ndConcerns* concerns.

We also show the tangling ratio metric, which is the ratio between the number of switch points and the woven code LOC. The higher the tangling ratio, the more the secondary code is tangled with the functional code. The lower the ratio, more localized the concern is.

The tangling ratio for all the examples is low, since the presented aspects where specialized for each application and require few join points. Therefore, the aspects are more localized to specific points. The aspect bloat, on the other hand, depicts some effectiveness of the aspects over the applied concern. In the first example, the LOC required to code the concern is the same as the aspect LOC, giving a positive efficiency to this aspect. On *LatnrmSpecialization*, the aspect requires more LOC than the actual concern, since we need to remove some code before applying the changes. The last two aspects show a reduced effectiveness of the aspects, since these two aspects are specific for their applications.

The measures obtained depict not only promising values but also some disapproving values. Separating the secondary concerns from the original application is a worth method for better modularization. However, if the aspect is not implemented for reusability, its effectiveness is reduced. The first aspect in Table II is more reusable than the others, since the strategy can be applied to different applications. In contrast, the *LatnrmSpecialization* is less reusable, since it is more specialized to the application.

V. Related Work

The following section describes an overview of the most relevant work about MATLAB compilation and AOP approaches. This related work was used by the authors as base for the development of the presented tool.

AspectJ[6] is an AOP extension for Java™ aimed at providing better modularity for developing Java programs. Developed by Xerox PARC and launched in August 1998, AspectJ contributes for a cleaner and better code by modularizing programs, providing solution for several CCCs such as monitoring, logging, debugging and synchronization. It is a general-purpose AOP language just like Java is a general-purpose OO language[11, 12]. AspectJ is faithful to its target language, Java, as it partially follows a similar structure of Java programs.

AspectMATLAB[7] provides AO extension for MATLAB [4] and was created thinking on the needs on developing scientific applications. This aspect language supports the same notions of other aspect oriented languages, but it gives them other terminologies: the pointcuts are called patterns; and advice is called a named action. This choice of terminology was based to clarify the user which is the matching objective and which is the action to take place on the matching pattern [4]. It is a language based on the aspect-oriented language AspectJ, previously described, adding some distinctive features that are imperative for MATLAB programmers, such as: the ability to capture multidimensional array accesses and loops on MATLAB; and to exposure the join point shadow information and combine it with the action over that join point, in other words, use the join point information to recreate the action [13]. The compiler for AspectMATLAB was based on the *abc* compiler [8] for AspectJ, and was built as an extension for the Natlab front-end [14], extending its rules to include aspects as a program entity.

AMADEUS[8] is a research project partially funded by FCT[9]. The AMADEUS project has the objective to enhance MATLAB development and implementation systems by using aspects. Specifically, the AMADEUS project focuses on optimizing programs in MATLAB by extending them with an AOP approach. This approach uses AOP to specify additional information such as type definition, to try different approaches to functions and variables, and to configure low-level representation of variables and expressions [15-17].

VI. Conclusions

This paper depicts a new aspect-oriented approach to the MATLAB language named MatlabWeaver. The new approach allows a vast set of join points and provides not only the common insert action but also different weaving actions, such

[6] Available at http://www.eclipse.org/aspectj/

[7] Available at http://www.sable.mcgill.ca/mclab/aspectmatlab/index.html

[8] Available at http://www.fe.up.pt/~specs/projects/AMADEUS

[9] Fundação para a Ciência e a Tecnologia: http://www.fct.mctes.pt

as type definition and variables default value. We described the use of MatlabWeaver in the context of different MATLAB applications for monitoring, specialization and type definition.

The case studies depict different strategies, evidencing the usefulness of MatlabWeaver when developing MATLAB applications that require essays to try different implementations and to separate secondary concerns that are not intended to be converted into the target language. The measurements exposed that the separation of secondary concerns into aspects significantly reduces the original code LOC, allowing easier understanding and maintainability of the main code and easier design of different implementations. However, when the aspects are more specialized for an application, its reusability is reduced. This makes the aspect impracticable in other applications.

We foresee the development of new actions for the weaver, such as sub-expression replacement and additional join point attributes to be defined (for instance, name replacement and statement order). Future plans also include a dynamic weaving, an important approach when some information can only be accessed in runtime or to try a different approach in runtime, without the constraint of terminating the execution to apply the different concern.

ACKNOWLEDGMENT

The authors acknowledge the support of the FP7 EU-funded project REFLECT and of the REFLECT members. We also acknowledge the previous contributions done in the context of the AMADEUS project.

REFERENCES

[1] G. Kiczales, "Aspect-oriented programming," ACM Computing Surveys (CSUR), vol. 28, p. 154, December 1996.

[2] A.-O. S. A. O. Website. (2011, 13th February). Aspect-Oriented Software Development. Available: http://www.aosd.net/

[3] T. Elrad, R. E. Filman, and A. Bader, "Aspect-oriented programming: Introduction," Communications of the ACM, vol. 44, pp. 29-32, 2001.

[4] T. Aslam, J. Doherty, A. Dubrau, and L. Hendren, "AspectMatlab: an aspect-oriented scientific programming language," presented at the Proceedings of the 9th International Conference on Aspect-Oriented Software Development, Rennes and Saint-Malo, France, 2010.

[5] J. M. P. Cardoso, J. Bispo, P. Pinto, R. Nobre, T. Carvalho, and P. Diniz, "The MATISSE MATLAB Compiler: A MATrix(MATLAB)-aware compiler InfraStructure for embedded computing SystEms, Technical Report, ICT-2009-4 REFLECT Project," December 2012.

[6] J. M. P. Cardoso, J. G. F. Coutinho, and T. Carvalho, "LARA Programming Language Specification, v2.0," REFLECT Internal Technical ReportSeptember 2012.

[7] J. G. d. F. Coutinho, T. Carvalho, S. Durand, J. M. P. Cardoso, R. Nobre, P. C. Diniz, et al., "Experiments with the LARA Aspect-Oriented Approach," presented at the International Conference on Aspect-Oriented Software Development (AOSD'12), Potsdam, Germany, 2012.

[8] J. G. F. Coutinho, T. Carvalho, S. Durand, and J. M. P. Cardoso, "The LARA Aspect-IR," REFLECT Internal Technical ReportMarch 2012.

[9] E. Balland, P. Brauner, R. Kopetz, P.-E. Moreau, and A. Reilles, "Tom: Piggybacking rewriting on java," presented at the 18th International Conference on Term Rewriting and Applications (RTA'07), Paris, France, 2007.

[10] E. Figueiredo, C. Sant'Anna, A. Garcia, T. T. Bartolomei, W. Cazzola, and A. Marchetto, "On the maintainability of aspect-oriented software: A concern-oriented measurement framework," in Software Maintenance and Reengineering, 2008. CSMR 2008. 12th European Conference on, 2008, pp. 183-192.

[11] B. Griswold, E. Hilsdale, J. Hugunin, W. Isberg, G. Kiczales, and M. Kersten, "Aspect-oriented programming with AspectJ," Copyright Xerox Corporation, vol. 2001, 1998.

[12] G. Kiczales, E. Hilsdale, J. Hugunin, M. Kersten, J. Palm, and W. Griswold, "An overview of AspectJ," ECOOP 2001—Object-Oriented Programming, pp. 327-354, 2001.

[13] T. E. F. Website. (2011, 27th January). AspectJ: crosscutting objects for better modularity. Available: http://www.eclipse.org/aspectj/

[14] J. M. P. Cardoso, T. Carvalho, J. G. d. F. Coutinho, W. Luk, R. Nobre, P. C. Diniz, et al., "LARA: An Aspect-Oriented Programming Language for Embedded Systems," presented at the International Conference on Aspect-Oriented Software Development (AOSD'12), Potsdam, Germany, 2012.

[15] R. Nobre, J. M. P. Cardoso, and P. C. Diniz, "Leveraging Type Knowledge for Efficient MATLAB to C Translation," Technical Report, Portugal2010.

[16] J. M. P. Cardoso, P. Diniz, M. P. Monteiro, J. M. Fernandes, and J. Saraiva, "A Domain-Specific Aspect Language for Transforming MATLAB Programs," presented at the Domain-Specific Aspect Language Workshop (DSAL'2010), part of the 9th International Conference on Aspect-Oriented Software Development (AOSD'2010), Rennes & Saint Malo, France, 2010.

[17] J. M. P. Cardoso, J. M. Fernandes, and M. Monteiro, "Adding Aspect-Oriented Features to MATLAB," presented at the SPLAT! 2006, Software Engineering Properties of Languages and Aspect Technologies, A workshop affiliated with AOSD 2006, Bonn, Germany, 2006.

SESSION 6

INFORMATION SYSTEMS AND INTEROPERABILITY

Eduardo Pinto
Architectural Key Dimensions for a Successful Electronic Health Records Implementation

Margarida Gomes
Policy debates in UK parliament: dataset, information retrieval and semantic web

Nelson Rodrigues
Towards Interoperability with Ontologies and Semantic Web Services in Manufacturing Domain

Architectural Key Dimensions for a Successful Electronic Health Records Implementation

Eduardo Pinto
Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
eduardo.pinto@fe.up.pt

***Abstract*—The availability of patient clinical data can be vital to a more effective diagnosis and treatment, by an healthcare professional. This information should be accessible regardless of context, place, time or where it was collected. In order to share this type of data, many countries have initiated projects aiming to implement Electronic Health Record (EHR) systems. Throughout the years, some were more successful than others but all of them were complex and difficult to materialise.**

The research involves the study of four international projects – in Canada, Denmark, England and France – launched with the goal of fostering the clinical data sharing in the respective countries, namely by implementing EHR-like systems. Those case studies served as data to identify the critical issues in this area.

To address the challenge of sharing clinical information, the authors believe to be necessary to act in three different dimensions of the problem: (1) the engagement of the stakeholders and the alignment of the system development with the business goals (2) the building of complex systems of systems with the capability to evolve and easily admit new peers (3) the interoperability between different systems which use different conventions and standards.

***Index Terms*—Electronic Health Record, Systems of systems, Software Architecture, Information Systems, Interoperability**

I. Introduction

Healthcare is one of the areas where new Information Technologies (IT) caused huge impact, helping to provide better services to the patients. Firstly, institutions implemented IT solutions to support their processes, allowing faster flow of the information but always in a very internal perspective. Throughout the years, the society's dynamic changed appearing an increasing necessity for sharing clinical information among different healthcare institutions [1]. This need represents enormous challenges both to the software development and business management areas.

Several projects were launched in many countries, such as in England or Canada for instance, with the goal of creating the well-know Electronic Health Record (EHR) systems. An EHR is a "longitudinal collection of electronic health information about individual patients and populations" [2]. The main objective is to provide clinical information about a patient where it needs to be consulted, independently of its origin or location, helping to avoid clinical errors or duplication of efforts and resources. It is supposed to be a mechanism for integrating healthcare information for the purpose of improving care quality [3].

Most of the times, these initiatives are not confined to the EHR concept but to the whole challenge of creating and maintaining accessible a set of clinical data of a patient, as complete as possible, independently of where was that information collected or who did it. Therefore, there might exist information sources other than the healthcare institutions information systems, like the patient himself, for instance. In this sense, the last years brought an emergent paradigm in which the patients have the central power over the information about themselves as long as their involvement in the process also grows [4]. The Personal Health Records (PHR) – which are systems whereby individuals can access, manage and share their health information which can be access by other in a private, secure, and confidential environment [5] – are the best example of it. On the other hand, the excess of information might be a trouble in several different situations. For that reason, some countries adopted concepts like Patient Summary (PS) [6] one which stands as a set of information that allows an healthcare professional to have a quick and easy overview over a patient. The case studies that the authors describe next are usually related to some or all of this concepts.

This research intends to understand the extent of the challenge that constitutes the implementation of a system like an EHR as a facilitator of clinical data sharing among several institutions. However, the work identifies those main challenges from an architectural point-of-view. Along with that, the authors expose the current methodologies, models and technologies that better address those challenges. In order to do so, some international cases are reviewed and the reasons for success or fail are summed up.

The paper starts with a description of some international case studies and their experience of these implementations at Section II. From that analysis, the authors extract the main challenges in these projects and expose them at Section III. Finally, the paper's conclusions are presented at Section IV.

II. Related Work

Several initiatives to implement EHR projects have been created over the years in several countries. Some countries have achieved more than others, but the common experience says that it stands not only as a technological challenge, but way beyond that [7]–[10]. The Table I does a brief summary of the four countries presented, from an architectural perspective. There are four classification parameters: (1) strategy - classify the process of implementation in terms of direction, (2) architecture - understand where it stands between the two

TABLE I
BRIEF SUMMARY FROM AN ARCHITECTURAL PERSPECTIVE OF THE FOUR COUNTRIES

	Canada	Denmark	England	France
Strategy	Bottom-up	Bottom-up	Top-down	Big-bang
Architecture	Distributed: fetched in real-time	Distributed	Hybrid: PS held national; the rest held locally	Distributed: host-based, fetched in real-time
Communication	SOA	Message-oriented	SOA	Document exchange
Standards	DICOM; SNOMED CT; HL7 v3; ICD10-CA	EDI (internal); CEN; SNOMED CT; DICOM	HL7 CDA; SNOMED CT	HL7 CDA; IHE profiles; SNOMED CT

extremes (completely distributed or completed centralized), (3) communication - state the approach used to allow the share of information and (4) standards - some of the used standards in that country either nomenclatures or encoding ones. It is easy to understand the variety of solutions that is possible to adopt only by these four examples. The details of these four case studies are described next.

A. Canada

In Canada, its First Ministers also created an non-for-profit organization called Canada Health Infoway back in 2001. It was specifically created to accelerate the EHR systems' development process, promoting the adoption of standards that guide to communication facilitation between different healthcare organizations. The effort would exceed 1.5 billion dollars [11].

In order to guide the development of the systems in each different province, Infoway provided a national framework called EHR Blueprint. The EHR Blueprint was a set of principles, guides and components. It states "a comprehensive description of the components necessary for the interoperable EHR and describes, in broad terms, how the components are envisioned to work together" [12]. EHRS Blueprint advocated that the best method (at least for their reality) was the creation of a shared reference information source that is populated by several health-care organizations around Canada. It is populated with clinical relevant data and is maintained externally from every health-care organization (or Points of Service, as designated in EHRS Blueprint). The Points of Service (PoS) are able to reference or pull data from the shared repository.

To summarise, the EHR Blueprint clearly identified the following key elements of its architecture (see Fig. 1):

- Point of Service Applications (PoS) — these are software applications or information systems that provide clinical information to the EHR system, working as gateway for gathering critical patient data, absolutely fundamental to the system's purposes;
- EHR Data Repositories — sometimes there is relevant clinical information that is not available in the PoS applications despite of being very relevant in a clinical decision making context. Instead, it is usually available through other systems. In this sense, the PoS applications are also responsible for pushing the data into these EHR Data Repositories – that become responsible for storing it and keeping it available to the users that might need that. Four logical clinical domain repositories are identified by EHRS Blueprint: Shared Health Record, Drug Information, Diagnostic Imaging and Laboratory;
- Registry Services — the information linkers. They provide the patients identification, matching them with the required clinical data. In order to guarantee the match of required and retrieved information, these services offer: Client Registry, Provider Registry, Location Registry and Terminology Registry;
- Longitudinal Record Services (LRS) — as the EHRS Blueprint is based on distributed data repositories, these LRS are responsible to bring together data from different registries and sources and normalizing it for common understanding when it is needed;
- Health Information Access Layer (HIAL) — a single standardized way of sharing and retrieving data from EHRi.

Another concept is the 'EHR Infostructure' (EHRi) which stands as an instance of the EHR for each different Canada's jurisdiction. In fact, there is not one single EHR, there are several (EHRi) replicated through the country, with the same structure, that are responsible for interacting with the local entities (PoS). The EHRi information is stored as copy of the original one and the EHR Blueprint was built following an Services-Oriented Architecture. Also, there is no single 'home' for the patient's electronic health record. Actually, each jurisdiction's EHRi holds and owns the data generated in health services from that jurisdiction.

In this case, the critiques started to raise up when one of the Infoway's goals: "By 2010, 50 per cent of Canadians and, by 2016, 100 per cent will have their electronic health record available to their authorized health care professionals." was not met. However, it seems that the problem was the lower rate of adoption instead of technical problems. McGrail et al. [14] argue that "after billions of dollars and nearly a decade of work by Infoway and the provinces, Canada is well back of the electronic records pack" and that Canada's "laggard position is certainly not a question of software design or lack or access to user-friendly technologies". Moreover, a qualitative study [15] reported that "key stakeholders identified funding, national standards, patient registries and digital imaging as important achievements of the e-health plan". On the other hand, the same study pointed out some issues of lack of clinicians engagement, e-health policy and a focus on national perspective rather than the regional one. As long with that, six provincial auditors joined and produced a report raising the possibility of the programme to need more funds [16], [17].

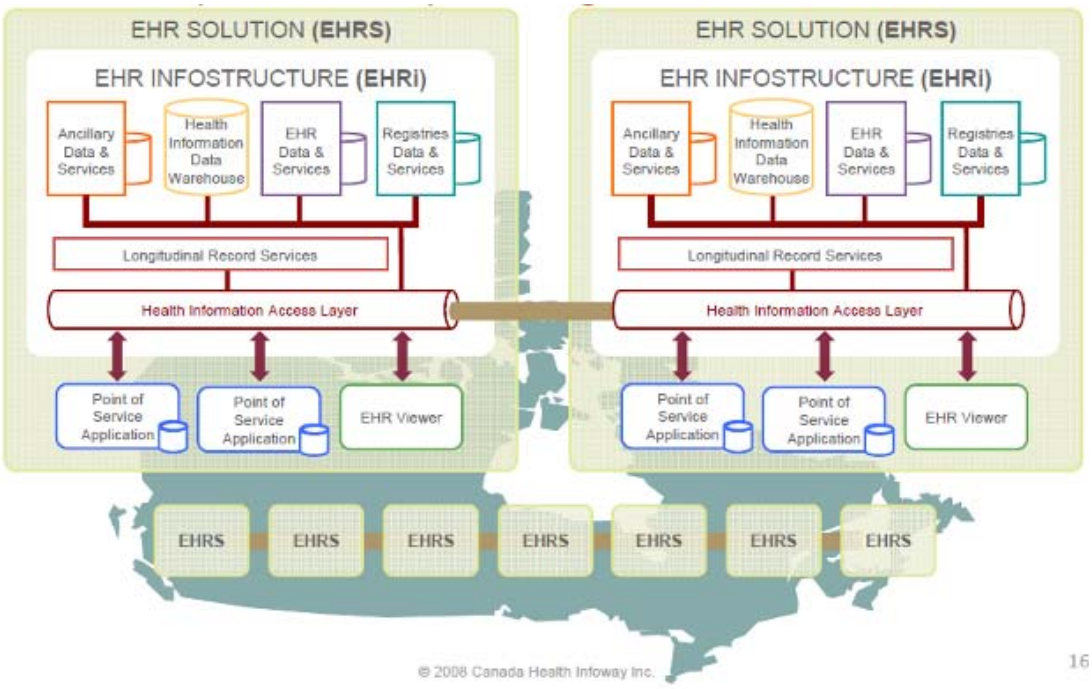


Fig. 1. The architecture general vision of the Canada EHR project [13]

More recently, Webster [18] advocates a change of strategy following British recommendations for not keep going with a top-down approach that is crafted with insufficient engagement of clinical users.

B. Denmark

Denmark is today recognized as one of the countries with higher success in the adoption of information technology in healthcare [9], [19]. At the end of 1994 the project organisation MedCom was created and mandated to establish a nationwide health care data network based on the Electronic Document Interchange (EDI) concept. The project was to be completed within a period of two years [20] and was the beginning of the Danish Health Data Network.

In Denmark, the doctors known as general practitioners serve as the "gatekeepers" of patients to other specialists and health professionals [21]. This context have enhanced the message-like solution that took place later. More than a national-wide solution, the programme was concerned about achieving value in the local regions by allowing fast and simple communication between the different stakeholders. The communication flow was based on six interactions, always with the general practitioner as a "nodal point" [22]: request laboratory results from hospital; prescription with pharmacies; reimbursement by assurance public health care; messages to emergency community care; radiology exams; referral and discharge information. This approach led to a "tremendous rise in messaging from 3 million per month in 2005 to 5 million per month in 2009 was much higher than expected" [19]. In that sense, for that past few years the Danes entities focus on moving from messaging to a web services approach, with the national databases being now used daily by physicians to look up lab results, patient identifiers, and medications.

Denmark has a National Patient Registry (DNPR) which has served as a data set of hospital contacts since 1977 [23]. Despite of being an administrative tool, the DNPR allowed the creation of a number of shared services, including the National Patient Index (NPI) that gathers all relevant information about individual patients. The NPI is Denmark's approach to the creation of a patient summary and the answer to the problem of inadequate access and overview of patient data as reports the European Commission [23].

Some authors argue that there was two key developments that led to the success of the Danish programme: "(1) creation of a nation-wide electronically accessible patient summary record, and (2) creation of a secure national health data network" [21]. Moreover, in 2006, Edwards [24] states some critical factors that led to success, such as: monetary incentives to the adoption of MedCom standards; precise standards worked out with the clinicians; gradual approach and realistic time frames; incentives to vendors; culture of consensus and a project-based approach.

In the late years, Denmark has shine with the national health network (MedCom) being "used by over 3/4 of the healthcare sector, altogether more than 5000 different organizations" [9]. Unlike other countries, when Denmark started to address these challenges, there was not a 10-years plan. In fact, it never existed. The strategy was always in perspective of 2 or 3 years and that seem to have worked pretty well.

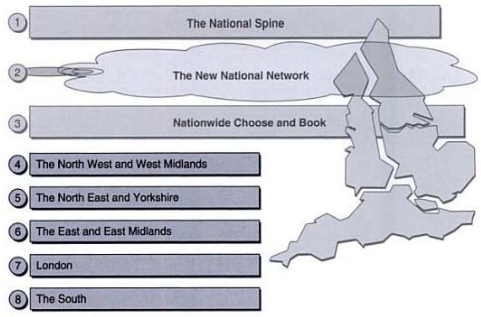


Fig. 2. The England's National Programme for IT strategic vision [25]

C. England

One of the most iconic projects took place in England, where the National Health Service National Programme for IT (NPfIT) was initiated in 2005. The NPfIT was an initiative to upgrade the National Health Service (NHS) to a centrally-mandated electronic health record for patients, connecting hundreds of hospitals and thousands of healthcare professionals and providing them relevant patient data by secure and certified means. The NPfIT has born as the world's largest civil information technology project, committing 12.4 billion pounds over 10 years in order to improve the quality of healthcare in England [25].

The NPfIT was made of eight specific-purpose systems (see Fig. 2): a big and national healthcare data repository, a national healthcare network, an electronic appointment booking system and five local service providers covering England's territory. It is important to notice that, in order to implement the five local clusters, five providers were contracted and made responsible for delivering the local services. The idea was that, in one hand, the providers would be challenged to compete between each other, speeding up the process of implementation. On the other hand, the risk would be lower since there was different suppliers implementing similar systems in parallel.

Concerning the interoperability issues, the program defined a set of standards, frameworks and implementation to guide and favour the interoperability between local systems and across them, called NHS Interoperability Toolkit (ITK). One of the key concepts was the use of a maturity-based approach, allowing the organizations to evolve through small steps, particularly for CDA documents. The step-by-step maturity model was meant to allow the organization to incrementally progress from sharing binary data to sharing fully-coded CDA documents.

The NHS Care Records was another of the NPfIT's components and two different types of records were projected:

- Summary Care Records — records held nationally. This concept, also known as Patient Summary, is meant to store essential information about a person, available in emergency situations, informing the health professionals about what medicines the patient is taking, the allergies that might suffer from or any known bad reactions to other medicines;
- Detailed Care Records — records held locally. This is a more comprehensive record which might store data from past exams and details, avoiding the necessity for repeating them, for example.

Since its beginning the project was the target of many critiques and source of many doubts. Back in 2004, John Powell advocated the importance of involving the clinicians in the changes as well as of showing the value to patient care [26]. Two years later, the problems seem to have grown more than the project itself, and to the complaints of lack of clinicians engagement were added statements of difficulty in deliver any value and doubt about the "fit for purpose" and created "clinical risks" as long as several major incident failures were succeeding [27]. A report by the King's Fund in 2007 also criticised the government's "apparent reluctance to audit and evaluate the programme", questioning their failure to develop a capable strategy [28]. In another tone, Sauer and Willcocks accused the funders to "pretend that they expect the promises to be delivered when they know they cannot be" and argued that was "time for game-playing to end and mature interaction to begin" [29]. Several authors written about the problems' causes and the things that could and should be learn [30]–[32], from lack of clinicians' engagement till wrong metrics, weak political leadership and one-size-fits-all solutions.

In September 2011, the NPfIT has been dismantled due the consecutive delays to deliver the fundamental services and the overrun of the programme's cost [33].

D. France

France is one of the countries with the best healthcare system in the world, having been classified inclusively as the "best health system in the world" by the World Health Organization (WHO) in June 2000. Following a set of initiatives in the late-nineties, in 2004 a law is approved to establish a Personal Medical Record (Dossier Médical Personnel - DMP) which aims to enable: (1) coordination of care (2) improve quality of care (3) continuity of care. However, that law was not only about the DMP, it had other directives to reform all the hospital sector that were "important enablers of healthcare delivery modernisation in France" [34]. In 2009, the "Rapport Fieschi" outlines semantic interoperability as the key goal and challenge of health information systems. At the same time, it is created the ASIP Sant, which incorporates GIP DMP (the Public Interest Group for Personal Medical Records).

The General Practitioner assumed a critical role in the DMP, being able to access it via the GP software or via the Internet but also accredited to transfer important documents into the DMP and hide document data upon request by the patient. Moreover, a number of patient-centred services are also envisaged [34]: consumer portal allowing the patient to access their healthcare record, to see the list of professionals who accessed their DMP and be informed of data updates; patient being able to manage the access rights of health professionals and update their personal information space as well as masking data.

In terms of the storage of the EHR, France is "the best example of a country that went with a host-based electronic health record system" [35]. The French patients are allowed to choose whichever data-host they want for their health record. Obviously, as prescribed by the French Decrees on Data Hosts, hosts have to be certified in order to be allowed to store clinical information. Another interesting fact is that, in France, an electronic health record can only be created after the consent of the patient ("opt-in" strategy).

In terms of standards, the French Dossier Medical Personnel adopted the IHE standards that were combined with the use of HL7 CDA for the Clinical Document Architecture.

III. Knowledge Domain Model

In this sense, it is easy to understand that the problem of sharing clinical information between different institutions contains several challenges. As we saw on the previous section, the problems that appear when implementing such a project are diverse and can have multiple sources and causes. The presented case studies show that the problems/challenges vary from high-level strategy to chosen standards. Despite of some of the issues discussed are not properly architectural problems, the fact is that all these problems are coupled, depend or affect some-how the architectural decisions. Therefore, the authors summarized the fundamental problems into three research questions:

- How to maintain the closeness between the system and the business goals guaranteeing the system's acceptance/adoption?
- How to build such complex and component-crowded systems in a way they would be able to change and evolve?
- How to make such different systems to speak and understand each other bringing forth patient care value?

A. Aligning the Business Goals with the System Development

Twenty years ago, the systems complexity was growing with an exponential velocity. However, most of the times, those systems were not able to fulfil the business needs and the problem was not lack of technology or knowledge but difficulties in understanding the business from those who were developing it. Thus, the software development was facing two problems at a time: in one hand, the systems were becoming huge and hugely complex; on the other hand, the systems were developed with few concerns about business orientation [36]. At that time, the concept of Enterprise Architecture appeared to address this problem. A lot of enterprise architecture models appeared and disappeared over the years. Three of the most known Enterprise Architecture frameworks are: The Zachman Framework, The Open Group Architecture Framework (TOGAF) and The Federal Enterprise Architecture Framework (FEAF). The Zachman Framework [37] aims to guarantee that all stakeholders' perspectives are being taken into account when developing a complex software system. In general terms, it is important to understand if all the artefacts are sufficiently focused and if the existing artefacts clarify all the players, from the business owner till the database designer, keeping all the visions aligned. First developed in 1995, TOGAF was based on the US Department of Defense Technical Architecture Framework for Information Management (TAFIM) [38]. TOGAF might be seen as a process for building an Enterprise Architecture. This framework states this building process as a continuous process of building multiple architectures from highly generic to highly specific ones, until reaching the organizational architecture level [36]. It provides an overall process template for architecture development activity and a narrative of each architecture phase, describing each one in terms of objectives, approach, inputs, steps and outputs to the architects [39]. Finnaly, the Federal Enterprise Architecture Framework appeared with the objective of serving as a platform for sharing processes, information and documentation among the U.S. Federal Agencies and other government agencies. This framework gathers two main characteristics of the two previous: in one hand, it defines a taxonomy – similar to the Zachman Framework – for artefacts classification; on the other hand, it suggests a process for building and implementing the architecture like TOGAF does.

B. Architecting Complex-Crowed Systems

Independently of the specificities of each country or region, the share of clinical information always depend on the capacity of a system to talk with the others systems. In this sense, the concept of Systems-of-Systems (SoS) is closely related to this idea of achieving value through the connecting and integration between several peer systems. A system is classified as an SoS when there is an "assemblage of several components which individually may be regarded as systems" and with two additional properties [40]: (1) operational independence of the components – each component is able to fully continue operating even when disassembled from the system (2) managerial independence of the components – the component systems do operate independently in fact. Maier suggests some architectural principles to deal with this kind of system like stable intermediate forms, policy triage, leverage at the interfaces and ensuring cooperation.

In addition to the obvious dependence of several systems, the fact of being a complex and large kind of project also deserves special attention and methodologies. In this field, the Ultra-Large-Scale Systems [41] and the Large-scale complex IT systems [42] are two terms that refer to the same type of systems. Independently of the designation used, these are systems identified by extreme size in every imaginable dimension: lines of code, stakeholders, number of systems, etc. Northrop, et al. state some characteristics as decentralized control,"unknowable" and conflicting requirements, continuous evolution and others [41].

If we lower the abstraction level, it would make sense to refer some architecture styles like Service-Oriented Architectures (SOA) and Resource-Oriented Architectures (ROA). SOA is an "architectural style that emphasizes implementation of components as modular services that can be discovered and used by clients" and that "emphasis on loose coupling between

interacting services" [43]. On the other hand, the ROA defines an easy access to the entities as well as the way that access should be done [44].

C. Transforming Data into Information, Knowledge and Value

Several registries are created and updated all over the healthcare institutions containing clinical data stored at divers information systems. The process of transforming that data into information which might be read by a professional is an essential step. Plus, the share of that information creating the knowledge necessary to offer better healthcare services is the ultimate challenge before creating real value to the patient.

The pursuit of interoperability is not possible without a clear definition of common languages and communication channels, usually called standards. In this sense, multiple standards can be found and compared [45], [46]. One of the most known organizations in this area is Health Level Seven International (HL7), which produced HL7 Messaging Standard. The Integrating Healthcare Enterprise (IHE) Profiles are other example, in this case these profiles define the systems involved (i.e., actors), the specific standards used, and the details needed to implement the solution. Also, the Digital Imaging and Communications in Medicine (DICOM) is an worldwide used standard for medical image communication, providing data structures and services and allowing the exchange of medical images and related information. Another standard is openEHR which develops specifications for implementing full EHR systems, pronouncing more in persistence as opposed to messaging, with the goal of achieve lifelong, patient centred, secure and shareable EHR.

The standards referred above are more about how to transfer the information. Another underlying question is the codification of that clinical information. In that case, the research is about terminologies and ontologies: Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), International Classification of Diseases (ICD) or Logical Observation Identifiers Names and Codes (LOINC).

D. The Peer-Generated Value

Despite the healthcare arena is a old-fashion one, it is interesting to understand how the new models of development could fit it. As known, the businesses are evolving from a product-oriented perspective to a relation-oriented one. This change of paradigm brought the client to the middle of the business, helping to create value. The Metropolis Model [47] appears as an attempt of describing really huge complex systems built from two basilar concepts: Open-source Software (OSS) and Community-Based Service Systems (CBSS). The Metropolis Model presents a new unified vision between the CBSS and the OSS, focusing deliberately in the crowd value generation through the definition of two levels: the kernel services and the periphery services.

In the healthcare, there is also a vast community ready to produce value and to be involved in the improvements that need to be done. The question is if the systems will be able to support and incentive that contribute.

IV. Conclusions

The implementation process of an EHR or a project that aims to foster clinical data sharing in a country is a challenge that congregates wide issues, from strategy and business processes to technology and interoperability problems. Despite existing the necessary technology to improve clinical processes and take the most advantage of IT, most country healthcare systems might not seem ready to fully exploit all its benefits, either because their institutions use paper-based processes yet, or their systems use specific standards created for their own use, or because the strategy adopted was simply not good enough to engage the stakeholders and make them the critical force pushing forward, etc.

None of the exposed EHR project was an invention or implied "scientific discovers", as the difference between what was being done was mostly in terms of scale. That is, there were already systems of systems implemented, there were already successful health information system implementations, standards and conventions had already been used to connect different health institutions, etc. The problem is when it is necessary to gather all this knowledge and apply it in a much bigger scale. On the other hand, if we say that the problem is the scale, so why not simply use divide and conquer strategy, making each part of it less complex? Denmark seems to successfully followed a strategy close to that, but is it all that simple? Will that strategy work in other country, with other culture and other health system as well as it worked in Denmark?

The authors believe that this projects have three critical dimensions: (1) keep the stakeholders (clinicians, nurses, institutions, etc) engaged as well as guaranteeing the system to be aligned with the business goals, using for example some principles of the Enterprise Architectures (2) adopt an agile architecture that is able to gathers new peers but also to allow the system to evolve and get better, following researches done in the area of System-of-Systems and Software Architecture (3) allow the institutions (peers) to effectively share clinical data creating value to the patient care services, taking part of the several standards and conventions existing in this area. It is not about technology that need to be invented, it is about how one is able to bring all this knowledge together and deliver not the "biggest computer program in the world" but the most valuable one.

References

[1] R. Haux, "Health information systems - past, present, future." *International journal of medical informatics*, vol. 75, no. 3-4, pp. 268–81, 2006.

[2] T. Gunter and N. Terry, "The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions," *Journal of Medical Internet Research*, vol. 7, no. 1, 2005.

[3] P. Orszag, "Evidence on the Costs and Benefits of Health Information Technology," in *Testimony before Congress*, vol. 24, 2008.

[4] M. J. Ball, C. Smith, and R. S. Bakalar, "Personal health records: empowering consumers." *Journal of healthcare information management : JHIM*, vol. 21, no. 1, pp. 76–86, Jan. 2007.

[5] C. for Health, "The Personal Health Working Group Final," Markle Fundation, Tech. Rep., 2003.

[6] epSOS, "epSOS: Patient Summary," http://www.epsos.eu/epsos-services/patient-summary.html, [Accessed: 11/11/2012].

[7] G. F. Anderson, B. K. Frogner, R. a. Johns, and U. E. Reinhardt, "Health care spending and use of information technology in OECD countries." *Health affairs (Project Hope)*, vol. 25, no. 3, pp. 819–31, 2006.

[8] A. K. Jha, D. Doolan, D. Grandt, T. Scott, and D. W. Bates, "The use of health information technology in seven nations." *International journal of medical informatics*, vol. 77, no. 12, pp. 848–54, Dec. 2008.

[9] D. Protti, I. Johansen, and F. Perez-Torres, "Comparing the application of Health Information Technology in primary care in Denmark and Andalucía, Spain." *International journal of medical informatics*, vol. 78, no. 4, pp. 270–83, Apr. 2009.

[10] B. H. Gray, T. Bowden, I. Johansen, and S. Koch, "Issues in International Health Policy Perspective on Meaningful Use ," vol. 28, no. November, 2011.

[11] C. H. Infoway, "Building a Healthy Legacy Together - Annual Report 2008/2009," Tech. Rep., 2009.

[12] ——, "EHRS Blueprint v2 Executive Overview," no. April, pp. 1–33, 2006.

[13] ——, "Model Report." [Online]. Available: https://knowledge.infoway-inforoute.ca/EHRSRA/index.html

[14] K. McGrail, M. Law, and P. C. Hébert, "No more dithering on e-health: let's keep patients safe instead." *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, vol. 182, no. 6, p. 535, Apr. 2010.

[15] R. Rozenblum, Y. Jang, E. Zimlichman, C. Salzberg, M. Tamblyn, D. Buckeridge, A. Forster, D. W. Bates, and R. Tamblyn, "A qualitative study of Canada's experience with the implementation of electronic health information technology," *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, vol. 183, no. 5, pp. 281–288, 2011.

[16] L. A. Offices, "Electronic Health Records in Canada - An overview of federal and provincial audit reports," Tech. Rep., 2010.

[17] P. C. Webster, "National electronic health records initiative remains muddled, auditors say." *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, vol. 182, no. 9, pp. E383–4, Jun. 2010.

[18] ——, "Centralized, nationwide electronic health records schemes under assault." *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, vol. 183, no. 15, pp. E1105–6, Oct. 2011.

[19] D. Protti and I. Johansen, "Widespread adoption of information technology in primary care physician offices in Denmark: a case study." *Issue brief (Commonwealth Fund)*, vol. 80, no. March, 2010.

[20] T. D. M. of Health, "A Danish Health Care Data Network In Two Years," 1996.

[21] M.-h. Kuo, A. Kushniruk, and E. Borycki, "Advances in Electronic Health Records in Denmark : From National Strategy to Effective Healthcare System Implementation," in *EFMI Special Topic Conference in Reykjavik*, 2010, pp. 96–99.

[22] T. D. M. of Health, "MedCom - the Danish Health Care Data Network towards the year 2000," Tech. Rep. March 1998, 1998.

[23] P. Doupi, E. Renko, S. Giest, and J. Dumortier, "Country Brief: Denmark," *Health (San Francisco)*, no. October, 2010.

[24] J. Edwards, "Case Study : Denmark's Achievements With Healthcare Information Exchange," *Gartner Industry Research Publication*, no. May, 2006.

[25] S. Brennan, *The NHS IT project: the biggest computer programme in the world - ever!* Radcliffe Publishing Ltd; 1 edition (April 2005), 2005.

[26] J. Powell, "NHS national programme for information technology: changes must involve clinicians and show the value to patient care," *BMJ: British Medical Journal*, vol. 328, no. 7449, p. 1200, May 2004.

[27] E. Wilkinson, "Is the UK health service IT project just too ambitious?" *The Lancet*, vol. 368, no. 9544, pp. 1317–1318, 2006.

[28] D. Wanless, J. Appleby, and A. Harrison, "Our future health secured," *A review of NHS funding*, 2007.

[29] C. Sauer and L. Willcocks, "Unreasonable expectations NHS IT, Greek choruses and the games institutions play around mega-programmes," *Journal of Information Technology*, vol. 22, no. 3, pp. 195–201, Sep. 2007.

[30] E. W. Coiera, "Lessons from the NHS National Programme for IT." *The Medical journal of Australia*, vol. 186, no. 1, pp. 3–4, Jan. 2007.

[31] C. Clegg and C. Shepherd, "The biggest computer programme in the world…ever!: time for a change in mindset?" *Journal of Information Technology*, vol. 22, no. 3, pp. 212–221, Jul. 2007.

[32] S. Brennan, "The National Programme for IT (NPfIT): is there a better way?" *Integrating Healthcare with Information and Communications Technology*, p. 95, 2009.

[33] D. of Health, "Dismantling the NHS National Programme for IT," http://mediacentre.dh.gov.uk/2011/09/22/dismantling-the-nhs-national-programme-for-it/, [Accessed: 14/03/2012].

[34] J. Artmann and S. Giest, "Country Brief: France," no. October, 2010.

[35] K. Stroetmann, J. Artmann, and V. N. Stroetmann, "European countries on their journey towards national eHealth infrastructures," *Final European …*, no. January, 2011.

[36] R. Sessions, "A Comparison of the Top Four Enterprise-Architecture Methodologies," pp. 1–28, 2007. [Online]. Available: http://msdn.microsoft.com/en-us/library/bb466232.aspx

[37] J. a. Zachman, "A framework for information systems architecture," *IBM Systems Journal*, vol. 26, no. 3, pp. 276–292, 1987.

[38] A. T. O. G. Josey, "TOGAF Version 9.1 Enterprise Edition," *Group*, pp. 1–13, 2011.

[39] S. Leist and G. Zellner, "Evaluation of current architecture frameworks," *Proceedings of the 2006 ACM symposium on Applied computing SAC 06*, p. 1546, 2006.

[40] M. Maier, "Architecting principles for systems-of-systems," *Systems Engineering*, 1998.

[41] L. Northrop, P. Feiler, R. Gabriel, J. Goodenough, R. Linger, T. Longstaff, R. Kazman, M. Klein, D. Schmidt, K. Sullivan, and K. Wallnau, *Ultra-Large-Scale Systems: The Software Challenge of the Future*. Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, 2006.

[42] I. Sommerville, D. Cliff, and R. Calinescu, "Large-scale complex IT systems," *Communications of the ACM*, vol. 55, no. 7, pp. 71–77, 2012.

[43] L. Srinivasan and J. Treadwell, "An overview of service-oriented architecture, web services and grid computing," *HP Software Global Business Unit*, vol. 2, 2005.

[44] H. Overdick, "The Resource-Oriented Architecture," *2007 IEEE Congress on Services (Services 2007)*, pp. 340–347, Jul. 2007.

[45] K. Atalag, D. Kingsford, C. Paton, and J. Warren, "Putting Health Record Interoperability Standards to Work," *electronic Journal of Health Informatics*, vol. 5, no. 1, pp. 1–17, 2010.

[46] M. Eichelberg, T. Aden, J. Riesmeier, A. Dogac, and G. Laleci, "A survey and analysis of Electronic Healthcare Record standards," *ACM Computing Surveys (CSUR)*, vol. 37, no. 4, pp. 277–315, Dec. 2005.

[47] R. Kazman and H.-M. Chen, "The metropolis model a new logic for development of crowdsourced systems," *Communications of the ACM*, vol. 52, no. 7, p. 76, Jul. 2009.

Policy debates in UK parliament: dataset, information retrieval and semantic web

Margarida Igreja Gomes
Faculty of Engineering
University of Porto
Rua Dr. Roberto Frias
4200-465 Porto, Portugal
pro12002@fe.up.pt

***Abstract*—In this paper is described the work done in the areas of information retrieval and ontology, whose motivation is the existence of large information repositories of very diverse nature and requirements for their organization, description, storage and research. This work included the construction of a parser to treat the dataset related to recent debates in the english parliament and the identification of users and their needs. A policy ontology, named twfy.OWL, was designed, represented and explored. The main classes, subclasses, properties and the data types were described. To accomplish the main representative interrogations a web interface was created. This simple, user-friendly, efficient and effective interface allows searching debates, scroll through the results and see specific debate details. The tools Solr, Protégé, Jena and Sesame resulted in a good combination for creating this search application used in the policy domain.**

I. Introduction

This study aims choosing, preparing and characterizing a dataset, applying a research tool, operating with free-text queries, designing, representating and explorating an ontology domain.

The work was organised in three phases. The the first phase, concerning the data preparation, consisted in researching repositories that provide datasets, choosing a domain and a set of data, exploratoring data analysis, characterizating the dataset, identifying its main properties and tasks to do researching on the data. In the second phase, that was related with information retrievel, included choosing a tool to use, analysing documentation and identifying its indexable components, using an Application Programming Interface (API) tool for generating indexes and setting the response to the questions, demonstrating the process of interrogation and information retrieval on the collection and evaluating the used technologies. The final phase, devoted to ontologies, consisted in analysing the chosen domain and its key concepts, evaluating the existing vocabularies for that domain, identifying features of the tool to use, constructing the ontology and operating the ontology through representative queries.

At this point the main definitions are presented for a better understanding of this paper. The concept of semantic web is described, by Tim Berners Lee, as an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. [1]

Another major concept is ontology, that was introduced by T. Gruber, as being an explicit specification of a conceptualization. [2]

Apart from the introduction, this paper presents, on section II, the dataset, describing the domain, exploratory data analysis, dataset characterization and a parser explanation and the Extensible Markup Language (XML) structure. On section III, information retrieval, starts with the users and their needs identification, analysis of an information retrieval tool and configuration and use of Solr. On section IV, semantic web, begins with a state of the art related with existing political ontologies, continues describing used ontology tools and technologies, then shows how the (They Work For You) TWFY ontology was constructed and presents some representative questions. On section V, an evaluation on the used tools and results is given.

Finally, the main conclusions of this work are pointed out and future potential work is identified.

II. Dataset

A. Domain

The chosen domain was the UK parliament and the selected dataset included two data lists, one containing the Members of Parliament (MP) and other the debates held in the House of Commons. [3]

This choice was due to several factors, among them, the quality, quantity and the accuracy of the data, the interest about english political organization and the data format available for download and use (API and files in Comma-Separated Values (CSV) format).

Regarding the authority source it comprises a group of twelve british citizens that prepare and publish this data on a voluntary basis.

All data have the english parliament as copyright holder (Parliamentary Copyright). As for the quality of the data, and after consulting the official website of the english parliament, it was found reliable.[4]

B. Exploratory data analysis

The data set chosen consists of a file of textual data (alphanumeric) CSV.

The data related with the MPs was downloaded, on 09/27/2012 at 19:27:59 with the size of 71KB.

Besides this method of collecting data the provided API was also used. However it was noticed that the source gives more details via API when comparing to the downloaded file.

To have access to this data via API it was necessary to make a registration in the TWFY website, which was followed by the assignment of a specific key (FK56abDtFG29DMwRLMD7GYZM). This way the data source may have some kind of control over the accessed data.

For example, the similar API to the MP file list, named getMP( ), returned the following additional fields: entered_house, left_house, entered_reason, left_reason, lastupdate, image, image_height, image_width, office and house.

Regarding the other list of data it resumes the debates held in the House of Commons, in a total of 2.234, between the 10th of January and the 9th of November of 2012.

C. Dataset characterization

In this section data is characterized and some of its properties are identified.

Making a first analysis of the data it was found that, in the list of MPs, there are 650 records containing the following six fields: personal ID, first name, last name, political party, constituency and Uniform Resource Identifier (URI).

After this quantification continued with some features that help a first approach to the data.

While the personal ID field is numeric, the others are textual. The ID field varies in the range 10.001 and 25.162. The party field can take a fixed set of values, among them, Labour, Liberal Democrat, Conservative, etc. The name field has titles associated, between them, Baroness, Viscount, Lord and Bishop. Finally, the MPs field is a URI that starts with the following predicate "http://www.theyworkforyou.com/mp/name".

After an initial processing of the data it was noticed that the character encoding is American National Standards Institute (ANSI). This feature was detected for the particular case of a registration with the name, Ynys Môn, when you convert the file "*. part" to "*. xlsx".

In order to obtain a more refined characterization of the data Google Refine tool was used. [5]

Begining using the API that returns the list of members of the english parliament and the option of Google Refine that lets entering the following Uniform Resource Locator (URL):

http://www.theyworkforyou.com/api/getMPs?key= FK56abDtFG29DMwRLMD7GYZM&output=xml

Then turned to the plugin reconciliation tool and used it to relate with the Freebase MPs data. [6]

The reconciliation was made at the field constituency with political districts (/ government / political_district) using the freebase electorate, first and last name.

The results obtained were as follows: 64% overlap, 36% to reconcile. Performing a manual reconciliation was obtained: 100% match, new 1%, 0% to reconcile.

Thus, it is seen that the data from Freebase is not as updated as data from TWFY website, with respect to members of parliament. After reconciliation with the political party (/ government / political_party) resulted in 30% coincide.

Then, a new column was created based on first and last names. The menu was clicked of the first name and added the selected expression value + " + cells ['Last name']. Value

From the reconciliation with political (/ government / politician) resulted 59% match, 0% new, 42% to reconcile. One could also add the attribute as a political party but the results were less good since only 1% match. Finally, before the manual reconciliation, 100% match, new 1%, 0% to reconcile were achieved.

In order to enrich these contents an extension of the data was done, that included, gender and date of birth of politicians.

So as to add columns from Freebase gender: female: 109, male: 448; empty: 92. The final results were then manually complemented, female: 145, male: 504. After insertion of gender for those lines it was not done any reconciliation.

In what respects to the date of birth results it was obtained 8 without results, 3 of which belong to MPs without topic. As many records have only the year of birth a column with this information was created using "Add column based on this column" and added the following: (value + -. ") Split (-") [0]

In order to introduce the summary table of UK parliament members, held with a Google Refine extension and data reconciliation between TWFY and Freebase, there is Figure 1.

In Figure 2 is presented one of the realized MPs data analyses. The graph shows the number of MPs ranked by political party (conservative, labour, etc.) and gender (female and male).

A political debate is characterized by an ID, unique and distinctive attribute, a title, a debate date and time, a debate location, speakers identification through an ID, a politician name, a party, their constituents and debate content.

D. Parser and XML structure

Once the data realted to the political debates were not available via API or a downloadable via file it was necessary to create a parser for the data housed on the TWFY website.

For this the Jsoup library was used. [7]

The problem was structured in several parts. One was the recovery of the debates URLs using the official website, other was the parser of each discussion and the last was storing the debate at an XML file.

In Code 1 is presented a portion of an original policy debate, before running the parser.

```
<debate id="2012-10-18a.564.0">
 <title>Backbench Business - 2nd Battalion
   the Royal Regiment of Fusiliers</title>
 <datetime>12:03 pm 2012-10-18</datetime>
 <place>UK Parliament</place>
 <speeches>
  <speech id="g516.1">
   <speaker id="40059">
    <name>John Baron</name>
    <party>Conservative</party>
    <constituency>Basildon and Billericay
    </constituency>
```

649 rows

Extensions: Freebase

Show as: rows records Show: 5 10 25 50 rows « first ‹ previous 61 - 70 next › last »

All	Person ID	Name	Date of birth	Year of birth	Gender	First name	Last name	Party	Constituency	URI
61.	24725	Karen Bradley Choose new match	1970-03-12	1970	Female Choose new match	Karen	Bradley	Conservative Party Choose new match	Staffordshire Moorlands Choose new match	http://www.theyworkforyou.co /mp/karen_bradley /staffordshire_moorlands
62.	10061	Ben Bradshaw Choose new match	1960-08-30	1960	Male Choose new match	Ben	Bradshaw	Labour Party Choose new match	Exeter Choose new match	http://www.theyworkforyou.co /mp/ben_bradshaw/exeter
63.	10062	Graham Brady Choose new match	1967-05-20	1967	Male Choose new match	Graham	Brady	Conservative Party Choose new match	Altrincham and Sale West Choose new match	http://www.theyworkforyou.co /mp/graham_brady /altrincham_and_sale_west
64.	10063	Tom Brake Choose new match	1962-05-06	1962	Male Choose new match	Tom	Brake	Liberal Democrats Choose new match	Carshalton and Wallington Choose new match	http://www.theyworkforyou.co /mp/tom_brake /carshalton_and_wallington
65.	24952	Angie Bray Choose new match	1953-10-13	1953	Female Choose new match	Angie	Bray	Conservative Party Choose new match	Ealing Central and Acton Choose new match	http://www.theyworkforyou.co /mp/angie_bray /ealing_central_and_acton

Fig. 1. *UK MPs characterization using Google Refine tool*

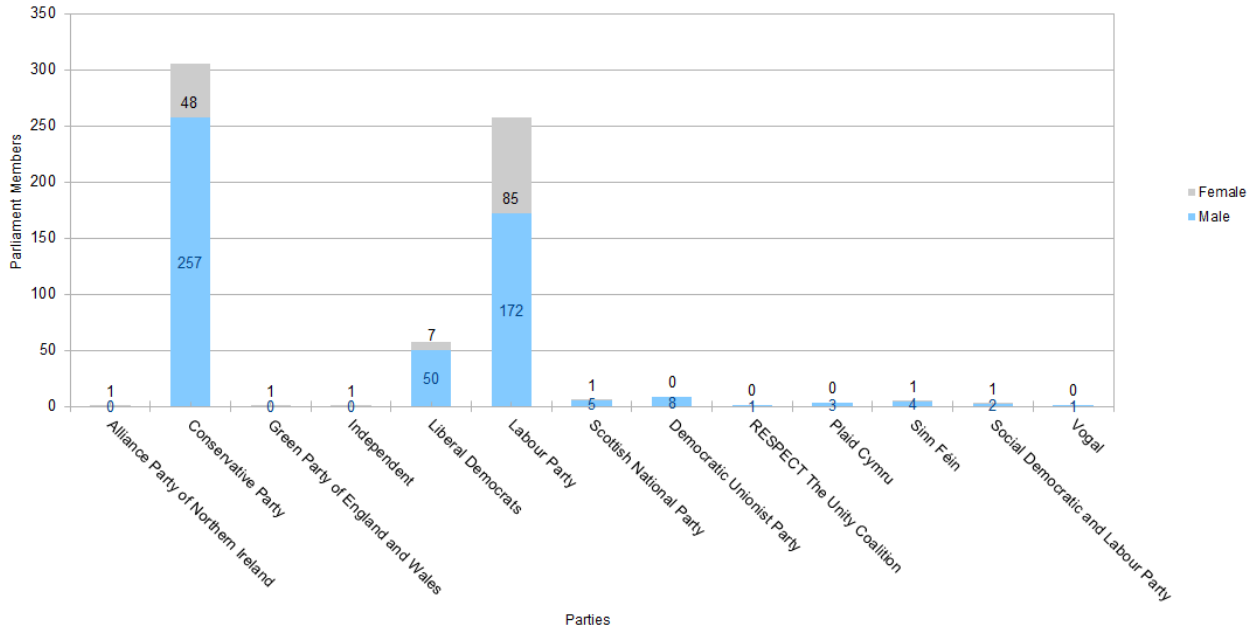


Fig. 2. *UK MPs sorted by party and gender*

```
  </speaker>
  <content>
   I beg to move,
   [...]
  </content>
 </speech>
 <speech id="g516.2">...</speech>
</speeches>
</debate>
```

Code 1. UK Policy debate (before parsing)

The new XML structure for the discussions is shown in Code 2.

It can be seen that each *doc* element is the speech of a debate with the information on the speech and the debate. The *add* element may have one or more *doc* elements.

```
<add>
 <doc>
  <field name="id">g6.4</field>
  <field name="debate-id">
    2012-01-10b.6.3
  </field>
  <field name="debate-title">Access to
   Drugs</field>
  <field name="debate-start">
    2012-30-10T02:30:00Z</field>
  <field name="speaker-id">40540</field>
  <field name="speaker-name">Stephen
    Metcalfe</field>
  <field name="speaker-party">
    Conservative</field>
  <field name="speaker-constituency">
    South  Basildon</field>
  <field name="content">What steps he is
    taking to ensure drugs  [...]
```

```
  </field>
  </doc>
  <doc>[...]</doc>
</add>
```

Code 2. UK Policy debate (after parsing)

III. Information retrieval

A. Users and their needs identification

Throughout this study there arose several possible questions to make the data set.

Thus, journalists and statesmen might need to know:

- Who is the most interventionist speaker?
- What is the debate with more interventions?
- What are the debates that took place between two dates?

Moreover, in the case of politics, they might be interested in knowing:

- What is a pary position concerning a specific theme?
- What were the interventions of a determined speake?

Finally, citizens might be curious about:

- In which debates was discussed a particular topic?

The previous list of questions could grow significantly given the wide range of themes presented in the data set. The relevance of the questions were filtered, as well as its viability given chosen tool characteristics - Solr.

B. Solr - Information retrieval tool

Among the available information retrieval tools (Lucene, Solr, Terrier, etc) Solr was selected.

This choice took into account the comparative study of several of these open source tools. [8]

Solr tool was selected because it has all the features of Lucene (since it is based on it) and allows incremental indexing (what does not happen with Terrier).

For this work it was used 3.6.1 of Apache Solr version, which is written in Java and requires the 1.5 Java or higher and an application server (such as Tomcat) with support for the 2.4 Servlet and provides an example with a Jetty server.

C. Using Solr

With the perspective of minimizing work an available example was used when instaling Solr. Thus it was necessary to change the debate XML schema.

First Sor was downloaded and unzipped. Then this tool tutorial was followed. [9] The steps of starting, indexing and asking questions were done. For the indexation of the parliament UK debates, it was proceed as follows on Code 3.

```
cd ${pasta-solr}\example
java -jar start.jar
cd exampledocs
java -jar post.jar debates/*.xml
```

Code 3. Indexing debates using Solr tool

After conducting the tool settings continued using the administration panel interface to raise questions.

IV. Semantic Web

A. Existing ontologies

After researching about the existing ontologies, which met policy domain and concepts were found the following repositories.

In the vast source of interesting DBpedia two ontologies were identified, one related to political parties and another associated with political debates. [10]

In a review of the literature it was found a reference to a political ontology, designated Government.owl, expressed in DARPA Agent Markup Language (DAML), characterized by having 84 classes, 78 object properties, data properties 3 and 53 subclasses.[11]

Given the information provided it was chosen not to reuse any of the existing ontologies since they were very general and a more specific ontology was needed. A a new ontology was created and was called twfy.owl.

B. Ontology tools and technologies

Given the vast array of available tools for working with ontologies the more complete, easier to use and providing better results was chosen. Thus, we turn to the Protégé, Jena and Sesame.

- Protégé
  Protégé is an open source environment, regardless of platform, for creating and editing ontologies and knowledge bases.
  It is Java-based, extensible and provides an environment for plug-and-play which makes it a flexible base for rapid prototyping and application development.[12]
  The Protégé platform allows two ways of modeling ontologies, one through the Protégé-Frames and another from the Protégé-OWL editors.[13]
- Jena
  The Jena is a project created by a core of researchers in the Semantic Web from HP.
  Its goal was to provide a platform in Java language that supported the use of Semantic Web for all applications that can be used.[14]
- Sesame
  Sesame was initially developed by the project OnTo-Knowledge, but together with businesses and Aidministrator OntoText, and presents a persistence layer for ontologies and inference.
  The code is Java, enabling easy migration for most operating systems.
  The development and use license is Lesser General Public License (LGPL). The system architecture is modular. There is a core module for encapsulating storage, called Repository Abstract Layer (RAL), functional modules for extraction, security, and management of data queries, and separately there are interfaces to access these modules. [15]
  The instalation of the Tomcat web application server was necessary. After the opnerdf sesame.war-and-openrdf

workbench.war was paleced in the / webapps, Tomcat was deployed. The repository was created in openrdf-workbench as a native RDF Schema database. After the repository, it was needed to add the ontology. To do this ”Add” in the ”Modify” tab was selected.

- SPARQL
  The SPARQL Protocol and RDF Query Language (SPARQL) is a W3C Recommendation since january 2008. Its main goal is to allow RDF files to be consulted through a SQL-like language.
  Allows the user to combine data from RDF files from different sources. The SPARQL is a data oriented language, or retrieves data files stored in RDF.
  It is built on the triplet pattern, ie: subject, predicate and object, and follows the same structure of building an RDF file. [16]

The entire application was developed using the Java NetBeans environment. Protégé was used for generating the ontology and creating the classes hierarchy and properties, Jena was used for populating the ontology, that is to say, generation individuals or instances and Sesame was used as the SPARQL endpoint, that is, as a search interface to use the ontology.

C. Ontology construction

In order to illustrate the TWFY ontology a graphical representation is shown in Figure 3.

Here it is identifiable:

- 6 classes (MP, Party, Politician, Debate, Speech and Person);
- 2 subclasses (MP is subclass of Politician, that is a subclass of Person);
- 3 inverse properties (isMemberOf - hasMember, inDebate - inSpeech and speaker - spoke);
- 2 different data types (dateTime and string);
- various functional properties.

For namespaces were used the following:

- Web Ontology Language (OWL), for classes, subclasses, functional and inverse properties;
- RDFS, for domain and range;
- XML Schema Definition (XSD), for data types.

To build this ontology it was necessary to create an hierarchy of classes and properties, with the ontologies editing tool Protégé , presented at Figure 4.

Then an individual named John Baron MP was created, who represents Basildon and Billericay constituency, has the identifier 40059, a member of the conservative party and made the speech g516.2.

In order to generate many instances easily and quickly Jena library was used. Regarding debates, for example, were established in 2.234 individuals.

D. Search engine

The search interface is very simple. There’s a search field at the beginning, where users can enter words to filter submitted

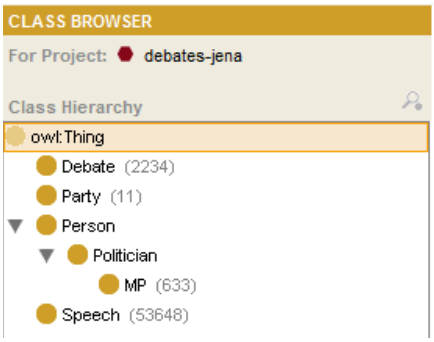


Fig. 4. *Class hierarchy of TWFY ontology using Protégé tool*

results. And a search button at the end. In the middle there are 5 radio buttons:

- All ontology triplets
- Members of Parliament names
- All MPs properties (constituency, name, etc.)
- Debate titles
- Debate contents

The result is shown with the words entered by the user highlighted and when there are more than 10 results, paging appears.

Figure 5 shows the layout of the described Graphical User Interface (GUI) with the question: which are the Members of the UK Parliament called ”smith”?

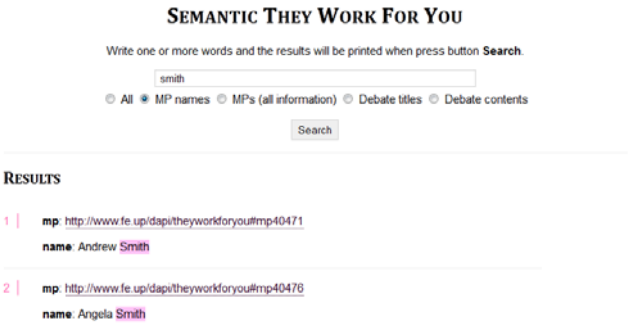


Fig. 5. *Search Interface (question and results) using Sesame tool*

In response two results were found, one named ”Andrew Smith” and another called ”Angela Smith”, sorted alphabetically.

V. Evaluation

A qualitative evaluation of the used technologies was made, namely, Solr, Jena, Sesame and SPARQL.

In what concerns to Solr, during the various questions asked, the tool emerged some considerations relevant to help pre-evaluation. For instance, a search for the name of the speaker (John Healey) with and without quotes was done and found that the results are different. In the first case arise all speakers with one or both of the names while in the second one only appear with the two speakers names simultaneously.

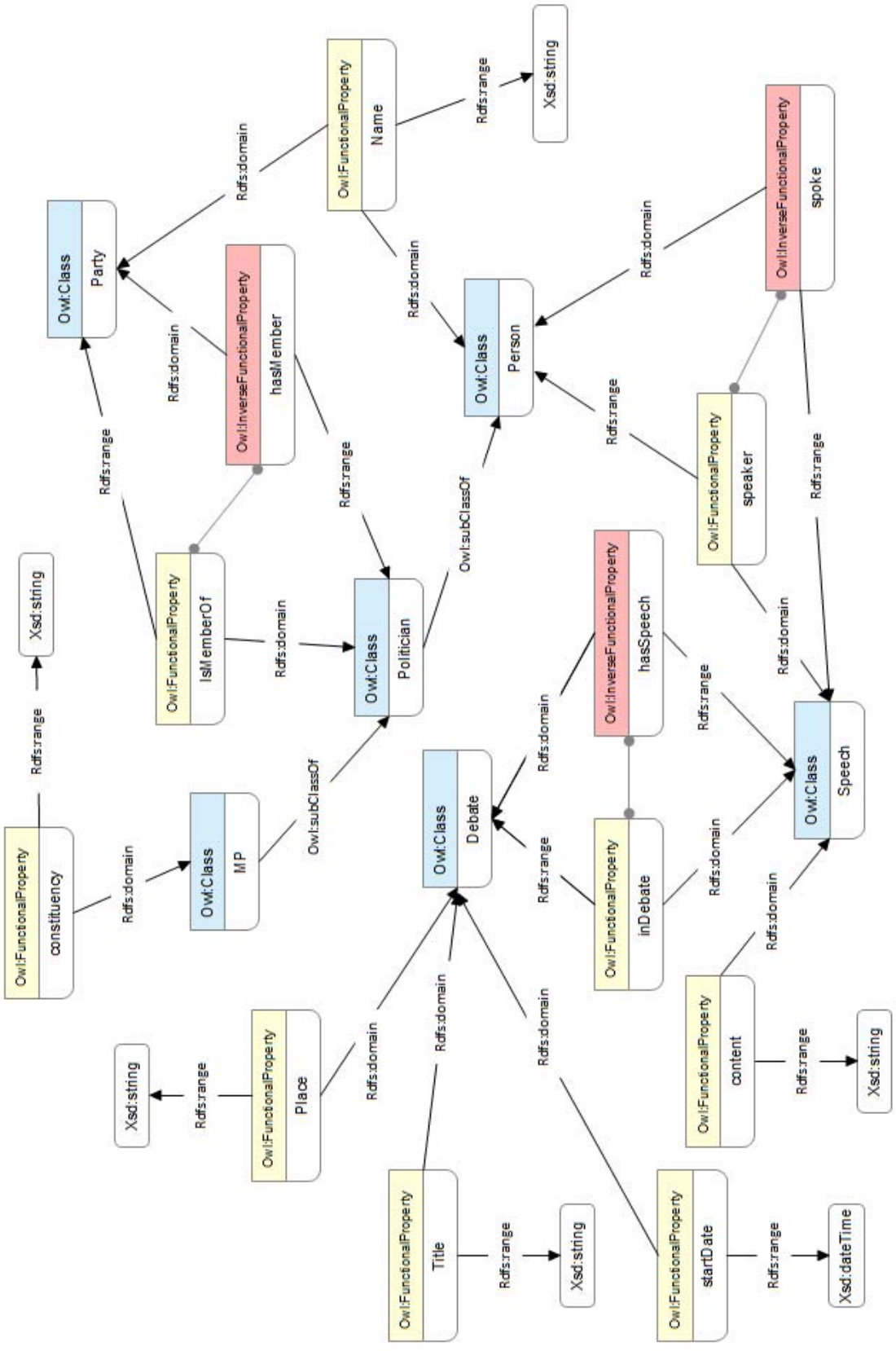


Fig. 3. *Graphical representation of TWFY ontology*

As to Jena, its learning curve was excellent. The basic features are very simple to understand and were sufficient to generate individuals. This was due largely to what is documented and its architecture.

With the Sesame tool it was possible to quickly create a SPARQL endpoint and has a compact API to facilitate the use of this computer-friendly interface. Its problem is the income, both using the Java Native system that stores files as a MySQL manager there are queries that are slow and consume too many resources.

As for the results it is needed to look at them from two perspectives: effectiveness and efficiency. The efficiency is very high. Not only the results that can be drawn directly from research, still achieve more complete results. But efficiency suffers as the price of great effectiveness. Simple queries involving few results are fast, but beyond them, consultations with many individuals, specifically the consultation on the debates contents, may take an average of 24 seconds. This is an unacceptable waiting time for an user and, therefore, should be improved.

VI. Conclusions

During this work it was recognized that our greatest effort was on analysing the data using a powerful open source tool, Google Refine, since, in the begining, it was chosen to use data that was available in CSV format and via API.

In order to complete this dataset it was decided to build an Hypertext Markup Language (HTML) parser to treat the data information related with the debates in the english parliament.

User requirements were defined and an useful application was conceived. Solr was used, a complete and flexible information retrieval tool that allows indexing documents and doing research.

A research was carried out to have the main references about the state art of the political ontologies and built up an ontology adapted to the root problem of the english parliament. Worked with different tools and libraries to design and make this ontology publicly available, thereby achieving one of the goals of the semantic web, data accessibility.

As future work would be very interesting to index the RDF graph and do research on it. Using an index that can improve the efficiency of some researches and in addition the results could be in an ordered ranking.

For this process could be used libraries as Lucene ARQ (LARQ) or Lucene Sail. Both based in Lucene but following different viewpoints. Lucene Sail is a cover of RDF storage that enables the use of Apache Lucene on the current storage. [17] LARQL is the integration of Lucene and ARQ, which is the search engine for Jena SPARQL. [18] Through this combination could obtain the benefits of indexing techniques, information retrieval and semantic web together.

Given the work done useful skills to prepare and retrieve information were acquired.

References

[1] T. Berners-Lee, J. Hendler and O. Lassila, *The Semantic Web*, Scientific American, 2001.

[2] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing", *International Journal of Human-Computer Studies*, Vol. 43, Issues 4-5, pp. 907-928, 1995.

[3] They Work For You website, Available: http://www.theyworkforyou.com, [Accessed: 09/2012].

[4] UK Parliament website, Available: http://www.parliament.uk, [Accessed: 09/2012].

[5] Google Refine documention, Available: http://code.google.com/p/google-refine/wiki/, [Accessed: 10/2012].

[6] Freebase Member of Parliament, Available: http://www.freebase.com/view/en/member_of_parliament, [Accessed: 10/2012].

[7] J. Hedley, "JAVA Parser for HTML, Available: http://jsoup.org, [Accessed: 10/2012].

[8] R. Baeza-Yates and C. Middleton, A comparison of open source search engines, Available: http://wrg.upf.edu/WRG/dctos/Middleton-Baeza.pdf, [Accessed: 11/2012].

[9] Solr Tutorial, The Apache Software Foundation, Available: http://lucene.apache.org/solr/api-3_6_1/doc-files/tutorial.html, [Accessed: 11/2012].

[10] DBpedia Political Party Ontology website, Available: http://live.dbpedia.org/ontology/PoliticalParty, [Accessed: 12/2012].

[11] A. Ortiz. "Polionto: Ontology reuse with automatic text extraction from political documents." *Proceedings of the 6th doctoral symposium in informatics engineering*, pp. 309-320, 2011.

[12] Protégé Tutorial, School of Medicine of Stanford University. Available: http://protege.stanford.edu/, [Accessed: 12/2012].

[13] Protégé Tutorial, School of Computer Science of Manchester, Available:http://owl.cs.manchester.ac.uk/tutorials/protegeowltutorial, [Accessed: 12/2012].

[14] Apache Jena Tutorial, Available: http://jena.apache.org, [Accessed: 12/2012].

[15] Sesame Tutorial, Available: http://www.openrdf.org, [Accessed: 12/2012].

[16] W3C Query language for RDF, Available: http://www.w3.org/TR/rdf-sparql-query, [Accessed: 12/2012].

[17] E. Minack, L. Sauermann, G. Grimnes, C. Fluit and J. Broekstra, The sesame lucenesail : Rdf queries with full-text search, 2008.

[18] Apache Software Fundation, Larq - adding free text searches to sparql, Available: http://jena.apache.org/documentation/larq/index.html, [Accessed: 12/2012].

Towards Interoperability with Ontologies and Semantic Web Services in Manufacturing Domain

Nelson Rodrigues[1,2]
[1] Faculty of Engineering of the University of Porto
Rua Dr. Roberto Frias, s/n 4200-465, Porto, Portugal
[2] Polytechnic Institute of Bragança, Campus Sta Apolonia, Apartado 1134,
5301-857 Bragança, Portugal
nrodrigues@ipb.pt

***Abstract*—Nowadays, companies are very dynamic, thus increasing their competitiveness is mandatory. This competitiveness is associated with dynamics and flexibility, a fast product changeover is needed. Interoperability has an important role to implement these features that companies need, in order to reduce time, effort and money. This paper describes how the production process can be improved with semantic models. With technical and methodological review through re-configuration low-level devices with Service Oriented Architecture and explores how Semantic Web Services can assist on this domain. The paper reviews the literature and points out the current research focusing Semantic Web Services and Ontologies applied to the manufacturing domain, and how interoperability in this field has proven to be essential.**

Keywords—Interoperability, Semantic Web Services, Ontologies, Manufacturing

I. INTRODUCTION

Currently, in the field of manufacturing, companies spend a lot of money and time installing new products and changes in production, such as configuration. It is necessary to adapt and optimize these processes to make them more dynamic in its reconfiguration. Nowadays, companies are very dynamic, thus increasing their competitiveness is mandatory. This competitiveness is associated with dynamics and flexibility, which are reached through solutions and interoperable infrastructures. Interoperability has an important role here and a huge impact, so creating dynamic, interoperable systems will bring several benefits, such as saving companies money, the ability to produce products in mass and quickly, and leaving an open door to, in the future, be able to integrate new processes without effort. *"it is estimated that 70% of the engineering teams' effort is involved in re-implementing the control"* [1].

In this way, technologies based on Services Oriented Architecture (SOA) become quite important to perform this passage. Collaboration between entities can be reflected in diverse domains. The current research direction is focused on developing semantic contracts among collaborating partners. Nevertheless, interoperability brings some problems, and it is necessary to identify them. There is the necessity to have a type of contract/meta-document among the entities, or different domains in the same company. This meta-document has the description and the semantics of the terms used. This problem can be solved with services and ontologies, which can clearly distinguish the semantics of terms. The Semantic Web Services is the connection of interoperable services with the semantics of the terms in a specific domain. With this expected dynamic it is necessary to re-think the companies IT architectures.

The remainder of this paper is as follows: Section II it briefly introduces the ontologies moreover their components and methodologies. Section III discusses the methodologies of SOA mechanisms. Section IV introduces the case study domain**.** Section V focuses on the engineering methodology on manufacturing according to Semantics Web Services. Finally, Section VI presents the paper results and section VII rounds up the paper with the conclusions.

II. ONTOLOGIES

The term "ontology" originates from the philosophy domain that has been adopted in Computer Silences, even though vague and not precise. Ontologies have been gradually used because of the need to represent knowledge in an area that has gained more interest in Semantic Web. Among the several definitions of ontology that can be found in the literature, the following one can be pointed out as the main definition:

An ontology "*is a formal, explicit specification of a shared conceptualization*". [2]

In face of this definition two different modelling layers can be described. The conceptualization level defines the concepts and relations among them, i.e., a way how to view a model from one perspective. The specification level specifies the conceptualization, in other words, is how formally (formal language) specifies how the world is seen.

In the computational world, ontologies are one way to describe, computationally, processable knowledge, but also to increase communication between computers and humans. There are three main reasons for using ontologies [3]:

1. Assist in communication between humans and computers.
2. Achieve interoperability between software systems.

3. Help improve the quality of design and system architecture software.

To accomplish the previous reasons, ontologies are developed taking into consideration knowledge reuse, sharing:

A. *Components*

Basically the smallest ontology is defined by a triple, namely the subject, predicate and object or in other terms concepts, relations and attributes.

Concepts are expressions that indicate domain entities with a complex structure that can be defined in terms of classes or objects, e.g., *Product*: (is a finished or semi-finished entity that is produced by the enterprise in a value-adding process). *Relations* or predicates establish the relationships among the concepts, e.g., *hasOperation (x, y): (*process plan x contains operation y). *Attributes* are values relative to properties of concepts, e.g., productID: a non-negative integer number that provides the unique identification of the product. *Restrictions* are conditions that should be satisfied when instantiating a class. Restrictions can be applied to the predicates, defining the range, domain and cardinality of the classes involved in the relation; and to the attributes of one class, defining the range and domain. In Fig 1. the ontology components are described, as well as how they are related.

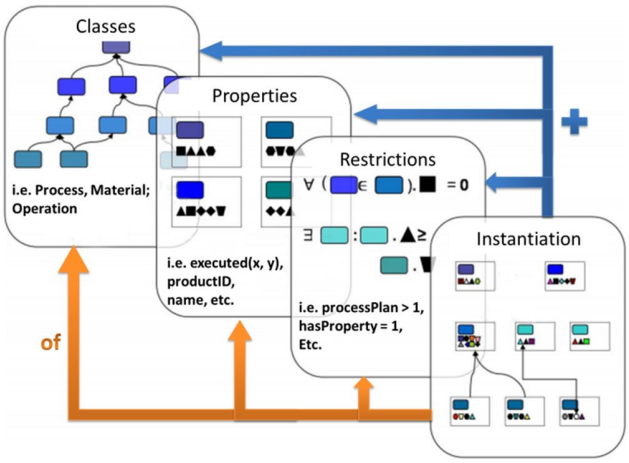


Fig.1. Ontology representation levels [47]

In the next sections, the several ontological components will be analysed. A crucial point is how to represent the ontology knowledge, i.e., the definition of classes, properties and relationships among classes. For this purpose a methodology should be followed.

B. *Methodology*

In literature it is possible to find particular frameworks that describe the stages step-by-step. Noy and McGuinness propose a methodology for the development of ontologies [4], for instance, Gruber proposed some principles [5]; the terms used in the ontology must be clear; the ontology should avoid doubts and misunderstandings about the terms used; the ontology design should support an easy expansion; among others.

The main idea in the development process of an ontology is to verify if existing ontologies can accomplish the proposed requirements, aiming to reuse ontologies; if the requirements are not accomplished, the option is to move to the next phases as illustrated in Fig 2.

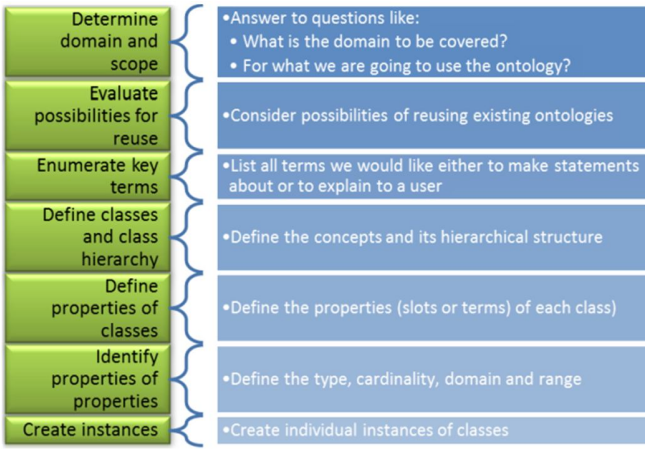


Fig.2. Methodology to build proposed ontologies [4].

As the ontology development evolves, it should exist a need for continuous evaluation of the ontology. At the end of a good agreement between the domain expert, users, and ontologies engineers, the ontology is concrete for that domain and generic to future improvements.

C. *Ontology Languages*

Nowadays, there are several languages to describe ontologies, giving here more attention to the more recent ones.

Resource Description Framework (RDF) [6] is one language used to develop ontologies based on the markup languages, e.g., the SGML (Standard Generalized Markup Language) and the XML (eXtensible Markup Language) [7]. Since XML is a declarative language, being quite limited, RDF appears to overcome these limitations, e.g., in terms of relations. RDF is used for representing information about resources on the web, thus constituting a basic ontology language. In RDF, the statements used to describe resources are represented as triples, consisting of a subject, predicate and object, i.e., {S, P, O}. The RDF(S) (Resource Description Framework Schema) is a semantic extension of RDF, which allows describing taxonomies of classes and properties, supporting the demand for creating a schema. The Web Ontology Language (OWL) [8] is another markup language that semantically extends RDF and RDFS, it derives from the DAML + OIL (DARPA Agent Markup Language - Ontology Inference Layer) [9]. OWL has a rich set of modelling constructors, offering improved pre-defined

templates, e.g., supporting the inclusion of restrictions in the concepts and predicates. OWL has a reasoning layer that allows representing an ontology in a more expressive manner.

D. *Ontology Frameworks*

The development of ontologies is a complex task that requires the support of proper frameworks which assist the creation or manipulation of ontologies and are able to express ontologies in one of many ontology languages. The use of these tools may lead to a more productive task in the design of ontologies, supporting the concurrent work of the ontology engineers and the domain experts. Several frameworks are currently available, namely OntoEdit [10], WebODE [11], Protégé [12] and Hozo [13]. Even the Protégé API can be used just like an API (Application Programming Interface). This API is implemented in Java and is essentially the same as Protégé, only without the graphic component; this API is to be used in conjunction with JENA [14]. Jena is a common framework that can be used in several approaches. It can be used individually but, it is explicitly used as the basis of Protégé API.

E. *Ontologies for manufacturing*

Ontologies are used in several and divers domains. In this paper the field is limited to the manufacturing domain.

In EU FP7 GRACE (inteGration of pRocess and quAlity Control using multi-agEnt technology) project, an manufacturing ontology was developed [15] to handle the knowledge exchanged among the Multi-agent system [16]. The EU FP6 PABADIS'PROMISE project proposed a reference meta-ontology for manufacturing [17]. ADACOR (ADAptive holonic COntrol aRchitecture for distributed manufacturing systems) for the manufacturing control domain, which was formalized with the DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) language [18]. MASON (Manufacturing's Semantics Ontology) introduces an ontology, in order to unify the ontologies using cognitive architectures, leaving to an implementation of a generic manufacturing ontology [19]. Other attempts to establish generic manufacturing ontologies are the NIST's description of shop data model [20], the Automation Objects [21], OOONEIDA focusing on the infrastructure of automation components by applying the semantic web technologies [22], and TOVE (Toronto Virtual Enterprise Ontology) that describes an ontology for virtual enterprise modelling [23]. The ISO 15926 standard [24] aims to support the integration of industrial automation systems. The challenge in manufacturing capability modelling lays in developing conceptual capability models that characterize several features of manufacturing, in terms of abstraction as well as formalization of the model.

III. SOA

This section describes the methodologies in Services-Oriented-*. SOC (computing) and SOA increase dramatically the services interoperability applied at inter-enterprises or intra-enterprises layers. The key concepts about SOA, are integration and reuse. SOA became very popular due to its features, which are very easy to implement and expand.

A. *SOA Components and methodology*

Nowadays SOA is a very popular architecture, due to an excellent solution to the many challenges of the current business, namely: providing a large component of agility through a quick response, and adaptability to changes, allowing companies to save time and money.

In SOA, one of the features is to minimize the relation of dependencies. This stateless services need to be dynamic. SOA follows certain principles, such as interoperability among the systems, reuse, granularity, modularity and componentization. Also it offers several services as: standardized service contract, loose coupling, service abstraction, service reusability, service statelessness, among others. Commonly there are three actions associated to a SOA: discovery, request and response, as it is described in Fig. 3, where three major entities are illustrated as well their own actions. The first action is the registration process (step 1), where the provider registers the services that can be performed. A discovering process of finding the service that provides the functionality that is required to the discovery service is usually called UDDI (*Universal Description, Discovery and Integration*) (step 2).

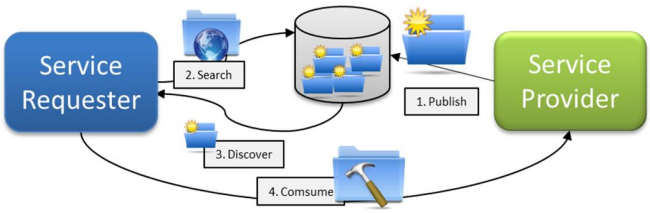


Fig.3. Service Oriented Architecture

The service provider receives from the UDDI Registry the interface needed (step 3) to invoke the service of the provider (step 4); the reply yields the output from the service (step 5).

B. *SOA Languages*

In the publish step, the protocol Web Services Description Language (WSDL) is used, in the following SOAP (Simple Object Access Protocol) messages are used, which are XML-based protocols that allows applications to exchange information. Also a XML-based protocol (the WSDL language) is responsible for describing Web services and how to access them, the definitions of ports, service name, operation, message, bindings and types. The message is well separated from its concrete instance, at the end is an interface description prepared to be reused. Thus an agreement, known as SLA (Service Level Agreement), is

necessary, responsible to handle the negotiated agreement between two entities; the service contract is then strictly defined. In what regards to the service contract, its anatomy is described in a WSDL, XML schema, and WS-Policy definition.

These concepts are well integrated due to the fact that there is good incorporation with collaborative automation, in the sense of self-governing, reusable and loosely-coupled distributed components. Also, due to this effort, modeling tools with Web Services protocols was developed to deal in a more abstract way, from the BPEL **(**Business Process Execution Language) [25] to the WSBPEL (Web Services Business Process Execution Language) [26], and WSFL (Web Services Flow Language) [27] proposed by IBM. The purpose is to assist on the modeling, which should be so abstract that if we put a new device and new processes in the automation system, they can integrate in order to achieve the objective: total integration. Some process can became automated to assist on the modelling, such as:

- **Orchestration** is an automatic and coordinated management of services taking into account a set of centralized services into a single one. In other words, consists on the combination of services to produce a more complex and useful services.
- **Choreography** describes each service as a service that knows exactly when to become active and with whom to interoperate, in a collaborative way. Both specifications should be implemented to make a more autonomous system.

C. *REST Web Services vs SOAP Web Services*

The major applications that fulfil the requisites of SOA use Web Services, which can be implemented in SOAP, REST and WSDL. However, there are several platforms to use/implement interoperability, e.g., REST, SOAP, XML-RPC, among others. Obviously there are some platforms that are more absorbed by the industry than others, perhaps because they bring short-term benefits, or because they are more easily implementable. REST is the simplest of all, being well regarded by the community for its simplicity. However the most significant difference is in relation to the interface definition, in RESTful systems is not necessary to describe one. This is the main reason why this implementation is simpler to use. Nevertheless there is some discrepancy regarding SOA implemented through REST, if it is, or not, a proper fulfilment of SOA; RESTful systems attempt to be implemented according SOA paradigm as seen in [28], it is alleged to be a Resource-Oriented Architecture (ROA) paradigm [29] and not SOA, where services are replaced by resources. One of the reasons of this paper is to identity which is the best platform, in a long term. REST technologies and SOAP, even if they are or not SOA compliant, both have their merits, but SOA becomes more semantic.

D. *SOA in manufacturing domain*

There is already some work in the manufacturing field, several European research project initiatives are available for consulting, based on the migration of industrial processes into service oriented architectures [30]: a FP7 project IMC-AESOP (www.imc-aesop.eu) [31], SOCRADES is a FP6 project addressing SOA-based in manufacturing (www.socrades.eu), focused on coupling web service enabled devices with enterprise applications [32]. Also in project SIRENA (www.sirena-itea.org), SOA is extended to a low-level domain such as embedded-devices (sensors and actuators) [33]. How to implement service-orientation through Multi-Agent Systems in industrial automation is described in [34]. A survey of the engineering of SOA is described in [35]. At [36] a practical example of device-level SOA is given.

IV. MANUFACTURING

Although companies realize the benefits of SOA implementation, they are still very apprehensive. Since usually when something is working well the main idea is to not change anything, but a restriction of evolution is placed every day on the company financial equation. As already demonstrated, the company must evolve. Then, why not implement it?

However industry has been slow when applying the agility and dynamic that SOA methodology offers, mainly because of the cost of replacing or develop from scratch their IT architecture, since most of the manufacturing companies have invested a considerable amount of money in manufacturing devices to handle the IT architecture. The massive computational power that has been developed in recent years is viewed as a disadvantage in addressing the problems in the companies today. Solutions were installed over time, of several application types, which enhance the automation or processing of each company's domain. Computational power is no longer so important. Nowadays the problematic is the integration/deployment of interoperable services. The developed solutions must assure trust and support along the time.

Thus, there is a problem in this temporal validation. After a system is developed and implemented in the production line, it takes years for refinement, validation and verification, thereby creating a dual problem. Firstly, the system must be generic as possible to leave open interoperability insights; secondly, the system must be specific enough to be able to accomplish the purpose for which it was developed. The process reengineering should be transparent, which it is not. In a long term, when a change is needed in a real model that has been used for years, according to the views of interoperability, it will be the most crucial test.

A. SOMAS

Manufacturing systems can be defined as "*a collection or arrangement of operations and process [...] to make (a) desired product(s) or component(s)*" [37]. To accomplish this concept a collaborative work must be performed between entities. The automation literature is replete of examples with Multi-Agent systems (MAS), which represent each entity in order to offer a solution to increase flexibility, distributed control, reduced complexity, etc. Currently trends in Service Oriented Multi-agent Systems (SOMAS) are being explored more often to increase interoperability, semantic descriptions, composition of services, among others, for further detail see [34] [35]. Additionally, some work related with agents based on ontology-based services to achieve interoperability is described in [46].

This way, such systems must assure modular capabilities, which mean that a system component can be divided into smaller components and mixed and matched in a variety of configurations. If the modularity is guaranteed by the system it is a good asset, and for the future one can implement new systems.

B. Manufacturing Standards

With the fast advance in technologies, the way how interoperable systems are developed should be rethought, because systems are developed taking into account present technology. In the future, technology will improve, and more systems and standards will appear, the question is how to create a system today that can be interoperable with the systems of yesterday and tomorrow. This is probably one of the major reasons why companies are so septic to implement such systems. The financial impact is very high to simple implement a system that only works in the manufacturing domain during one or two decades.

It is necessary to create rigid standards that assure this problem, in order to convince the companies that the system that they are using follows the standard points. The academic community is behind some standards to support and try to increase the implementation of SOA in the manufacturing domain. Also it is necessary to understand the industrial problems and solutions. On other hand, the industry must understand clearly the benefits of SOA and the openings that will appear with the academic research involvement, since industrial standards will be created.

These standards are also based on communication protocols, and to the message content specifications frameworks and architectures. However, there are many standards that already exist (as it will be seen next), but they do not consider the different knowledge entities. The problem with standards occurs when it is necessary to create two entities automatically interoperable; this approach is understandable in theory, but in practice it is very difficult to put two entities that were created from different approaches, and even without any of the methodologies, created without any standards.

V. Semantic Web Services & Ontologies

It is necessary to guarantee common understanding and data semantics among distributed entities (also reuse and share of knowledge). Ontologies can increase how the knowledge is expressed as it was seen in the first sections of this document. Moreover, ontologies can increase the semantic of the services, like the processes, how the knowledge is exchanged between two entities where both can understand the meaning [38]. It is mandatory to use a meta-model to exchange interoperability, and these Meta-Models can be made in several formats, namely in XML [7], RDF or OWL (Web Ontology Language) [8]. But in order to create interoperability this is not enough: it is necessary to implement services that express more than simple functions.

A. Semantic Web Services Motivation

In previous chapters it was mentioned that Web Services can offer features such as modular, self-describing, self-contained applications that are accessible over the Internet, being these, also, some of SOA characteristics.

Sometimes Web Services are confused with SOA, but SOA does not specifically mean Web Services. As an alternative, Web Services can be realized as a specialized SOA implementation. However, Web Services Description Language (WSDL) does not contain semantic descriptions of the operations, the notion is simple: join ontological notions in Web Services (WSDL). The combination of these two concepts makes Web Services more semantic, capable of expressing the semantics of their services. An evolution, taking into account this initiative, is illustrated in Fig. 4.

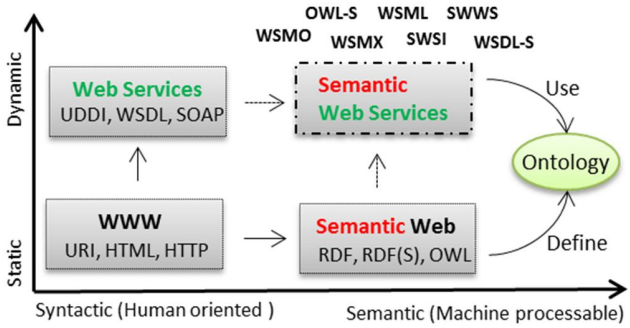


Fig.4. Evolution of the WEB. (adapted from [39])

The Web has experienced changes in its anatomy, becoming more dynamic and semantic.

B. Semantic Web Services Languages and Protocols

Approaches and initiatives which aim to specify Web Services using semantics and ontologies include: OWL-S [40], the SWSI, SWWS, WSML, WSMO and WSMX that can be view in more detail in [41]. WSDL-S [42] defines new elements and annotations for already existing elements, which offer a great potential to implement SWS.

C. Semantic Web Services in Manufacturing Domain

In the industry domain the Implementation of SOA can exist at different levels: on high-level and very similar to other companies (in the field of industry or not), or low-level, on devices of their own manufactures [42].

An IT infrastructure for the heterogeneous message communication available is Enterprise Service Bus (ESB) [43]. This infrastructure is probably the most used approach when a communication channel for SOA is needed. However this communication layer does not represent the meaning of each message. An FP6 project, SUPER (www.ip-super.org) [44], aims to achieve a higher degree of automation in discovery and mediation of co-operating services. The goal can be described by a "*Semantic Business Process Modeling"* [45], which is to follow a usage of semantic technologies, as Semantic Web Services, in the process modeling phase, creating a Semantic Service Bus (SSB) as an enhancement of the general ESB. Some projects offer semantically enhanced business process modeling and design of semantic ESB, such as: Object Management Group (www.omg.org), FP6 R&D project STASIS (www.stasis-project.net), and OPUCE (www.opuce.tid.es). The concept of SSB was also adopted in the SPIKE project (Secure Process-oriented Integrative Service Infrastructure for Networked Enterprises). The objectives of these projects are to recognize and provide observation in the manufacturing domain of the interoperable systems integration.

1) OWL-S

Mapping services and processes at a low-level domain in WSDL files, which describe the operations process is simple, the difficult part is to describe in semantics such services/operations in the WSDL file. One advantage of OWL-S is the specification of semantic concepts. OWL-S is represented by three main concepts (grounding, service and profile). It can automatically discover, invoke, compose and monitor resources, allowing then the offering of services. The challenge of integrating the approach involves two services that complement each other. The *ServiceProfiler* is responsible to fully describe the request service, namely what the service does. *ServiceGrounding* specifies how to use the service, in order to execute it. And *ServiceModel* gives information on how the service works.

2) WSDL-S

But WSDL-S can also be used to do this process. Having some benefits compared to the OWL-S, in particular the simplicity in the implementation transition and a wider range on what regards the types of modeling. WSDL-S is based on mapping annotations of a WSDL file, see Fig. 5 as an example. Therefore the selection process of services, discovering, description of services, invocation and composition, becomes more automatic and not dependent on the interpretation that each engineer intendeds to give, thus solving some terminology ambiguities that might exist.

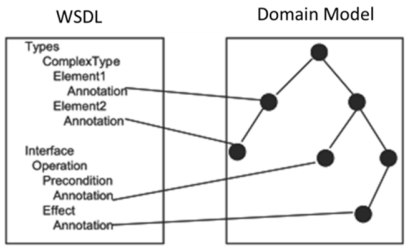


Fig.5. Association between WSDL and Ontologies from [42].

When using either WSDL-S or OWL-S it is necessary an automatic composition service. Web Services in WSDL can be matched against an announced OWL-S Web Service in an autonomous way [1]. A mediator is necessary to perform these and have the decision support centralized.

D. Practical example Mediators

To understand how the ontology concepts are associated with services and semantic used by the agents, let's get back to the example of SOA in the automation domain. One of the research paths that was followed, when placing in the services way the goals and functionalities, was to make this automatic switch, in order to achieve a machine-interpretable and human-interpretable transition, by defining the features and services of each device through Services notions. A Mediator should thus be used to aggregate services in SOA systems, applied to this domain.

The mediator is responsible for solving some mismatches in order to give to the systems the interoperability that they need. Another type of mediator is the "OO Mediator" that is responsible for mediating the ontologies, to merge, align and map, in order to retrieve integrated and homogeneous solutions. For example, if SOMAS is used, an agent can very easily provide a new service or a new process. In Fig. 6 the *Agent A* can easily reasoning that the service "*Dispatch pallet type B*" from the *Agent* B has similar features as the service "*Dispatch pallet A*", so if it is more rentable for divers variables to use different services, the agent can match, merge, discover, monitor or infer new services. The inference can be performed through the ontologies with the SWRL (Semantic Web Rule Language), where rules are used to assert a specific combination, e.g., the combination of the *hasParen*t and *hasBrother* properties implies the *hasUncle* property.

$$hasParent(?x1,?x2) \wedge hasBrother(?x2,?x3) \Rightarrow hasUncle(?x1,?x3)$$

In the manufacturing domain, some rules could be executed as some of them could be created implicitly during the process. Through SWS (Semantic Web Services) this step can be more easily automatized. The idea is to use a Mediator to take advantage of semantics. The composition of new services is according to the semantics of each one. Creating a semi-automatic composition performed by the *Mediator*.

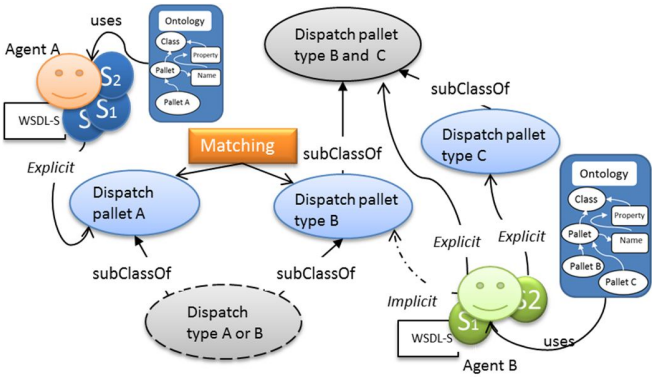


Fig.6. An example of reason matching services.

Centralized system has access to services in order to control the creation and integration of services.

E. *New knowledge for the services*

One very important point is the implementation of inference, know implementing the methods/techniques which will infer new facts or rules, making this inference an obvious reason in the context of this article, brings new relations and describes the best services. Thus the entities in charge of this service will emerge in new services, making them more semantic. To achieve that goal, some inference skills are necessary. Techniques asserted (asserted triples are those that were populated by triples from merging several sources) or inferred (are triples that were inferred by inference rules) can bring this type of new knowledge to create new or better services.

VI. Discussion

Implementing SOA architecture can be very difficult. As already seen, if well implemented it brings plenty of benefits, however if poorly implemented can harm a company financially and structurally. To achieve this, it is vital to follow some type of guidelines in order to not make mistakes. It is important to recognize the benefits, but just as important to know where the failures occur when doing this integration on a real company.

A. *Semantic benefits in Automation*

Semantic Web Services can be implemented in several domains, with several profits, but the key is highlighting the potential benefits of SWS in manufacturing. In an abstract way it is possible to simplify the development of flexible reducing development costs and time; create more robust systems; because is simpler, the software system maintenance will be easier as aggregation processes. In a low-level perspective allows assisting in automating service selection, fast reconfiguration, more agile automation, flexibility, without the need for system re-engineering.

B. *Before SOA, After SOA*

Right after implementation of a SOA system, it is expected to be more dynamic, to reuse and share services, more collaborative, very integrated and interoperable scenario. Adding SOA to the automation domain has clear benefits. Collaborate with industrial companies is mandatory to achieve a conclusion about SOA in industry. The industry, in order to be able and conscious that SOA brings benefits horizontally and vertically, must see the results, being these all about numbers and costs, and SOA can has a major influence in those results.

VII. Conclusions

This survey takes a trip along the current trends in manufacturing domain. By analysing the approaches of this paper, it is noticed that companies are ready to increase their responsiveness by changing from a static, inflexible and slow architecture to evolve into a dynamic, faster and agile one. The independencies in a typical operating model were tight coupled among systems of coordination and now, companies are improved to a loose coupling among systems of coordination.

Ontology and services can help on the heterogeneous conversations between the entities, creating then an interoperable system. SOA is perhaps the greatest revolution in business and industrial companies. That tends to link the functional processes of enterprises to the use of productive technologies. The sooner a company starts to use SOA, the sooner it will be ready to provide the best of services, thereby creating, in advance, more rivalry in relation to its competitors. In low-level SOA implementations, namely SOA in automation industry, it enables companies to perform an optimized business management, and a better final product allowing reconfiguration at real time. The engineers' efforts can be focus on dynamic systems in order to create such system that allows to infer new processes.

However this work is not finished, it is necessary to perform a validation. The interoperability of the system will be put into test when the re-engineering step arrives, so it is necessary to make sure that the system is interoperable. It is mandatory, in the future, the implementation of real scenarios to get real results, being a path to follow so that these approaches are absorbed by the manufacturing industry. Another path to follow is the large computational power spent to create this easy integration to be on a Cloud and profit from its benefits and reduced costs.

References

[1] A. W. Colombo, F. Jammes, H. Smit, R. Harrison, J. L. M. Lastra, and I. M. Delamer, "Service-oriented architectures for collaborative automation," in *Industrial Electronics Society, 2005. IECON 2005. 31st Annual Conference of IEEE*, 2005, p. 6 pp.

[2] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," *Int. J. Hum.-Comput. Stud.*, vol. 43, no. 5–6, pp. 907–928, 1995.

[3] M. Uschold, V. R. Benjamins, B. Ch, A. Gomez-perez, N. Guarino, and R. Jasper, "A Framework for Understanding and Classifying Ontology Applications," 1999.

[4] N. F. Noy and D. L. Mcguinness, "Ontology Development 101: A Guide to Creating Your First Ontology," 2001.

[5] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowl. Acquis.*, vol. 5, no. 2, 1993, pp. 199–220.

[6] O. Lassila, R. R. Swick, W. Wide, and W. Consortium, "Resource Description Framework (RDF) Model and Syntax Specification." 1998.

[7] W3C, "XML. Extensible markup lan guage (XML) 1.0." .

[8] W3C, "OWL. OWL Web ontology language overview," 2004. [Online]. Available: www.w3.org/TR/2004/REC-owl-features-20040210/.

[9] I. Horrocks, "DAML+OIL: a Description Logic for the Semantic Web," *IEEE Data Engineering Bulletin*, vol. 25, pp. 4–9, 2002.

[10] Y. Sure, M. Erdmann, J. Angele, S. Staab, R. Studer, and D. Wenke, "Ontoedit: Collaborative ontology development for the semantic web," 2002, pp. 221–235.

[11] Óscar Corcho, M. Fernéndez-López, A. Gómez-Pérez, and Óscar Vicente, "WebODE: An integrated workbench for ontology representation, reasoning, and exchange," in *IN: PROCEEDINGS OF EKAW 2002. LNCS 2473*, 2002, pp. 138–153.

[12] J. H. Gennari, M. A. Musen, R. W. Fergerson, W. E. Grosso, M. Crubzy, H. Eriksson, N. F. Noy, and S. W. Tu, "The evolution of Protégé: an environment for knowledge-based systems development," *Int. J. Hum.-Comput. Stud.*, vol. 58, no. 1, pp. 89–123, 2003.

[13] K. Kozaki, Y. Kitamura, M. Ikeda, and R. Mizoguchi, "Hozo: An Environment for Building/Using Ontologies Based on a Fundamental Consideration of Role and Relationship," in *Proc. of EKAW2002*, 2002, pp. 213–218.

[14] HP, "Jena - A Semantic Web Framework for Java." 2002.

[15] P. Leitao, N. Rodrigues, C. Turrin, A. Pagani, and P. Petrali, "GRACE ontology inteGrating pRocess and quAlity Control," in *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*, 2012, pp. 4348–4353.

[16] P. Leitao and N. Rodrigues, "Multi-agent system for on-demand production integrating production and quality control," in *Proceedings of the 5th international conference on Industrial applications of holonic and multi-agent systems for manufacturing*, 2011, pp. 84–93.

[17] L. Ferrarini, C. Veber, A. Luder, J. Peschke, A. Kalogeras, J. Gialelis, J. Rode, D. Wunsch, and V. Chapurlat, "Control Architecture for Reconfigurable Manufacturing Systems: the PABADIS'PROMISE approach," in *Emerging Technologies and Factory Automation, 2006. ETFA '06. IEEE Conference on*, 2006, pp. 545–552.

[18] S. Borgo and P. Leitao, "The Role of Foundational Ontologies in Manufacturing Domain Applications," in *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*, vol. 3290, R. Meersman and Z. Tari, Eds. Springer Berlin / Heidelberg, 2004, pp. 670–688.

[19] S. Lemaignan, A. Siadat, J.-Y. Dantan, and A. Semenenko, "MASON: A Proposal For An Ontology Of Manufacturing Domain," in *Distributed Intelligent Systems: Collective Intelligence and Its Applications, 2006. DIS 2006. IEEE Workshop on*, 2006, pp. 195–200.

[20] C. Mclean, Y. T. Lee, G. Shao, and F. Riddick, "Shop data model and interface specification," in *NISTIR 7198, National Institute of Standards and Technology*, 2005.

[21] O. J. L. Orozco and J. L. M. Lastra, "Using semantic web technologies to describe automation objects," *International Journal of Manufacturing Research*, vol. 1, no. 4, pp. 482–503, 2007.

[22] V. Vyatkin, J. Christensen, J. L. M. Lastra, and F. Auinger, "OOONEIDA: an open, object-oriented knowledge economy for intelligent distributed automation," in *Industrial Informatics, 2003. INDIN 2003. Proceedings. IEEE International Conference on*, 2003, pp. 79–88.

[23] M. S. Fox, "The TOVE Project Towards a Common-Sense Model of the Enterprise," in *Proceedings of the 5th international conference on Industrial and engineering applications of artificial intelligence and expert systems*, 1992, pp. 25–34.

[24] R. Batres, M. West, D. Leal, D. Price, K. Masaki, Y. Shimada, T. Fuchino, and Y. Naka, "An upper ontology based on ISO 15926," *Computers & Chemical Engineering*, vol. 31, no. 5–6, pp. 519–534, 2007.

[25] F. Curbera, R. Khalaf, N. Mukhi, S. Tai, and S. Weerawarana, "The next step in Web services," *Commun. ACM*, vol. 46, no. 10, pp. 29–34, 2003.

[26] M. Kloppmann, D. Koenig, F. Leymann, A. Rickayzen, C. von Riegen, P. Schmidt, and I. Trickovic, "WS-BPEL Extension for People - BPEL4People," 2005.

[27] F. Leymann, "WSFL. Web services flow language," 2001.

[28] T. Erl, B. Carlyle, C. Pautasso, and R. Balasubramanian, *SOA with REST: Principles, Patterns & Constraints for Building Enterprise Solutions with REST*. 2012.

[29] Leonard Richardson and Sam Ruby, *RESTful Web Services Web services for the real world*. O'Reilly Media, 2007, p. 454.

[30] T. B. Jerker Delsing, Fredrik Rosenqvist, Oscar Carlsson, Armando W. Colombo, "Migration of Industrial Process Control Systemsinto Service Oriented Architecture," in *IECON*, 2012.

[31] S. Karnouskos, A. W. Colombo, F. Jammes, J. Delsing, and T. Bangemann, "Towards an architecture for service-oriented process monitoring and control," in *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society*, 2010, pp. 1385–1391.

[32] A. Cannata, M. Gerosa, and M. Taisch, "SOCRADES: A framework for developing intelligent systems in manufacturing," in *Industrial Engineering and Engineering Management, 2008. IEEM 2008. IEEE International Conference on*, 2008, pp. 1904–1908.

[33] F. Jammes, H. Smit, J. L. M. Lastra, and I. M. Delamer, "Orchestration of service-oriented manufacturing processes," in *Emerging Technologies and Factory Automation, 2005. ETFA 2005. 10th IEEE Conference on*, 2005, vol. 1, p. 8 pp. –624.

[34] J. M. Mendes, F. Restivo, P. Leitão, and A. W. Colombo, "Injecting service-orientation into multi-agent systems in industrial automation," in *Proceedings of the 10th international conference on Artifical intelligence and soft computing: Part II*, 2010, pp. 313–320.

[35] J. M. Mendes, P. Leitão, F. Restivo, A. W. Colombo, and B. A, "Engineering of service-oriented automation systems: a survey," in *3rd I*PROMS Virtual International Conference on Innovative Production Machines and Systems*, 2007.

[36] L. M. S. De Souza, P. Spiess, D. Guinard, M. Köhler, S. Karnouskos, and D. Savio, "SOCRADES: a web service based shop floor integration infrastructure," in *Proceedings of the 1st international conference on The internet of things*, 2008, pp. 50–67.

[37] J. T. Black, *The Design of the Factory with a Future*. 1991.

[38] L. Obrst, "Ontologies for semantically interoperable systems," in *Proceedings of the twelfth international conference on Information and knowledge management*, 2003, pp. 366–369.

[39] F. Curbera, W. A. Nagy, and S. Weerawarana, "Web Services: Why and How," in *In OOPSLA 2001 Workshop on Object-Oriented Web Services. ACM*, 2001.

[40] *OWL-S. OWL-based Web service ontology*. 2004.

[41] J. Cardoso, *Semantic Web Services: Theory, Tools and Applications*. Hershey, PA, USA: IGI Publishing, 2007.

[42] R. Akkiraju, J. Farell, J. A. Miller, M. Nagarajan, A. Sheth, and K. Verma, "Web Service Semantics - (WSDL-S)," in *{W3C} Workshop on Frameworks for Semantics in Web Services*, 2005.

[43] D. Chappell, *Enterprise Service Bus: Theory in Practice*. 2004.

[44] D. Karastoyanova, B. Wetzstein, T. van Lessen, D. Wutke, J. Nitzsche, and F. Leymann, "Semantic Service Bus: Architecture and Implementation of a Next Generation Middleware," in *Proceedings of the 23rd International Conference on Data Engineering Workshops, ICDE 2007, 15-20 April 2007, Istanbul, Turkey*, 2007, pp. 347–354.

[45] M. Hepp, F. Leymann, J. Domingue, A. Wahler, E. Wahler, and D. Fensel, "Semantic Business Process Management: A Vision Towards Using Semantic Web Services for Business Process Management," in *In Proceedings of the IEEE ICEBE 2005*, 2005, pp. 535–540.

[46] A. Malucelli, "Ontology-based services for agents interoperability", Doctoral thesis, University of Porto, 2006.

[47] Z. Sun (2009) "Using ontology and semantic web services to support modeling in systems biology", Doctoral thesis, University College London, 2009.

PAPERS IN ALPHABETICAL ORDER

AUTHORS IN ALPHABETICAL ORDER

ISBN 978-972-752-151-7

9789727521517