Maria Luísa Lopes de Faria

A proposal for an integrated framework capable of aggregating IoT data with diverse data types

São Paulo
2017

Maria Luísa Lopes de Faria

A proposal for an integrated framework capable of aggregating IoT data with diverse data types

Tese apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Doutor em Ciências

Área de Concentração:
Sistemas Eletrônicos

Supervisor: Carlos Eduardo Cugnasca

São Paulo
2017

Catalogação-na-publicação

Lopes de Faria, Maria Luisa
A proposal for an integrated framework capable of aggregating IoT data with diverse data types / M. L. Lopes de Faria, C. E. Cugnasca -- São Paulo, 2016.
117 p.

Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Sistemas Eletrônicos.

1.Data Science for IoT 2.Smart lifestyle 3.Smart City I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Sistemas Eletrônicos II.t. III.Cugnasca, Carlos Eduardo

Acknowledgements

Uma garotinha nasceu num lar muito humilde. Filha de um laminador e de uma dona de casa. Cresceu amada pelos seus pais e foi filha única até os 6 anos de idade. Os pais incentivaram os seus estudos, mas podiam apenas enviá-la para a escola pública. Assim ela passou todo o primário e ginásio numa escola municipal. Já com 15 anos aquela garotinha transformou-se numa jovem que escolheu estudar numa escola técnica estadual e formou-se Técnica em Administração de Empresas. Todavia, um dos requisitos daquela escola era um estágio, e assim ela foi estagiar no Setor de Pesquisa e Desenvolvimento da Engenharia da Volkswagen. Influenciada por aquele ambiente, ela mudou sua área de estudos. Dali nasceu o sonho de ingressar num curso de Engenharia da USP. Ela passeou pelos corredores da Escola Politécnica sonhando em fazer parte daquela renomada instituição. Porém, prestou vestibular duas vezes, mas não atingiu a pontuação necessária para ingressar. Desapontada com seu desempenho, ela tentou estudar numa escola de Engenharia particular, mas sérias restrições financeiras a impediram de prosseguir com os estudos. Até que um dia um milagre aconteceu. Ao orar ela ouviu uma voz que lhe disse: "Vai para São Carlos!". Ela ficou confusa, não tinha informações sobre São Carlos, não tinha Internet. Dois anos depois prestou FUVEST e foi aprovada para o curso de Matemática na UFSCar. Assim, partiu para São Carlos e foi muito feliz durante todo o tempo de sua graduação. Deus a abençoou muito. Já formada pela UFSCar, ousou assistir algumas aulas no ITA como aluna especial. Desistiu do ITA, mas Deus a fortaleceu e assim ela se formou Mestre. O sonho de estudar na Escola Politécnica estava adormecido, mas nunca foi esquecido. Cinco anos depois, ela finalmente realizou aquele grande sonho. Ingressou no Doutorado da USP e seu primeiro professor foi o Prof. Carlos Cugnasca. Desde então, muitas vezes ela andou pelos corredores da Poli sem acreditar que finalmente era aluna da USP e da Escola Politécnica. Aquele sonho impossível havia se transformado numa realidade feliz! A USP lhe abriu as portas do mundo, propiciando-lhe um estágio no exterior e parte do seu doutorado na Inglaterra. Ela conheceu pessoas do mundo inteiro, mas nada alterou o orgulho que ela tem do seu Deus, do seu país, da USP e de seus pais. Durante o período de seu Doutorado conheceu muitas pessoas geniais, fez grandes amigos, ganhou duas irmãs Chinesas e foi professora voluntária de Português para estrangeiros. Hoje, aquela garotinha humilde está pulando de alegria dentro do meu coração! É com esta grande alegria que eu agradeço imensamente a TODOS que contribuiram para este final feliz! Que Deus lhes abençoe!

"O meu Deus vai na minha frente, Ele endireitará os caminhos tortos, Ele quebrará as portas de bronze e depedaçará os ferrolhos de ferro, Ele me dará os tesouros das escuridades e as riquezas encobertas!"

Deus é muito bom pra mim!

"Adaptação do verso de Isaias 45.2,3"

(Bíblia Sagrada)

Abstract

The volume of information in the Internet is growing exponentially. The ability to find intelligible information among vast amounts of data is transforming the human vision of the universe and everything within it. The underlying question then becomes which methods or techniques can be applied to transform the raw data into something intelligible, active and personal? This question is explored in this document by investigating techniques that improve intelligence for systems in order to make them perceptive/active to the recent information shared by each individual. Consequently, the main objective of this thesis is to enhance the experience of the user (individual) by providing a broad perspective about an event, which could result in improved ideas and better decisions. Therefore, three different data sources (individual data, sensor data, web data) have been investigated. This thesis includes research into techniques that process, interpret and reduce these data. By aggregating these techniques into a platform it is possible to deliver personalised information to applications and services. The contribution of this thesis is twofold. First, it presents a novel process that has shifted its focus from IoT technology to the user (or smart citizen). Second, this research shows that huge volumes of data can be reduced if the underlying sensor signal has adequate spectral properties to be filtered and good results can be obtained when employing a filtered sensor signal in applications. By investigating these areas it is possible to contribute to this new interconnected society by offering socially aware applications and services.

Keywords: Data Science for Internet of Things. Smart Lifestyle. Smart Cities.

List of Figures

Figure 1 – WoT encompasses data from several sources . . . 14
Figure 2 – The smart universe around the smart citizen. . . . 16
Figure 3 – Systems must be improved in order to understand not only the environment, but also the user and his sentiments. . . . 17
Figure 4 – Data fusion can provide better information, resulting in wise ideas and better decisions. . . . 18
Figure 5 – Turning Data into Wisdom . . . 20
Figure 6 – Scheme of organization. . . . 22
Figure 7 – The individual perspective is the target. Correlated data must be aggregated taking into consideration firstly individual data. . . . 24
Figure 8 – Data types noted by Mousannif & Khalil . . . 25
Figure 9 – An architecture using physical and virtual sensors . . . 31
Figure 10 – Chronology of Internet and Web evolution. . . . 32
Figure 11 – QA systems at the top of Web evolution. . . . 34
Figure 12 – Videk enriches sensor data with web data, however it does not uses citizen data. . . . 36
Figure 13 – Videk and its four main components. . . . 37
Figure 14 – Videk is based on multi-layered and scalable architecture. . . . 37
Figure 15 – Event calendar - system interface. . . . 38
Figure 16 – Open Data led by the government in the UK. . . . 39
Figure 17 – Three areas that must be considered when investigating emotions. . . . 42
Figure 18 – Steps to extract Sentiment Analyses. . . . 43
Figure 19 – Types of Data. . . . 47
Figure 20 – The quantity of unstructured data is growing rapidly. . . . 48
Figure 21 – Individual data is the main driver. Correlated data must be aggregated taking into consideration individual data first. . . . 50
Figure 22 – The position of the sensors 4, 16 and 26 are highlighted in green. . . . 54
Figure 23 – Question answering system design. . . . 57
Figure 24 – A vision that drives this research. . . . 58
Figure 25 – An example of a web dashboard report. . . . 59
Figure 26 – This figure represents a summary of Chapter 3. . . . 60
Figure 27 – The created .txt files were standardized as: @usuario: - tweets. . . . 61
Figure 28 – Each word has a numerical value (or weight) according to its frequency. 62
Figure 29 – Results presented by SOM algorithm. Each node has three numbers: the ID of the node, the number of iterations (or training) and the number of profiles in each module . . . 62

Figure 30 – Each word has a weight according to its frequency. 63
Figure 31 – The position of the sensors 4, 16 and 26 are highlighted in green. . . . 65
Figure 32 – Original signal received from the sensors aggregated in 4096 time bins. 66
Figure 33 – Periodogram. 66
Figure 34 – Vectors of coefficients concatenated. 67
Figure 35 – Cumulative Energy. 67
Figure 36 – The approximate signal has a high degree of similarity when compared with the original signal. 68
Figure 37 – Sensor 4 and Sensor 16 have different values of temperature. 70
Figure 38 – Sensor 4 & Sensor 16. A comparison between the error produced by the orginal signal and the error produced by the denoised signal. 71
Figure 39 – Sensor 4 & Sensor 16. A comparison between the error produced by the orginal signal and the error produced by the denoised signal. 72
Figure 40 – Sensor 4 & Sensor 16. A comparison between the error produced by the orginal signal and the error produced by the denoised signal. 73
Figure 41 – Noise can invade both the Wireless Sensor Network and the Internet. . 73
Figure 42 – Sensor 4 - Humidity Data - Autocorrelation function. 74
Figure 43 – Sensor 4 - Humidity Data - Partial Autocorrelation function. 75
Figure 44 – Periodogram. 75
Figure 45 – Variance. 76
Figure 46 – A simple QA system developed during WISS. 78
Figure 47 – Scientific name of the species and essential information about it. 81
Figure 48 – Data of three environments must be fused and comprehended. 82
Figure 49 – Wikitude is a type of navigation based on Augmented Reality. 83
Figure 50 – The Red Cross Museum suggests a hashtag and offers interaction through Instagram. 84
Figure 51 – A general overview of the idea. 93
Figure 52 – Design of the proposed digital tool. 96
Figure 53 – Design of the proposed mobile app to the citizen. 97

List of Tables

Table 1 – Original Data Set 68
Table 2 – DWT coefficients necessary to reconstruct the signal 68
Table 3 – Main data collected, sensors used, communication technology, and information provided for the main stakeholders in our proposal 95

List of abbreviations and acronyms

ACF	Autocorrelation Function
API	Application Programming Interface
AR	Augmented Reality
BER	Bit Error Rate
CoreNLP	A set of natural language analysis tools
DBMS	Database Management System
DBN	Deep Belief Networks
DBPedia	DB for Database, aiming to extract structured information from Wikipedia making this information available on the Web.
DWT	Discrete Wavelet Transforms
EPC	Electronic Product Code
GPS	Global Positioning System
GUI	Graphical User Interface-
HTTP	Hypertext Transfer Protocol
IEML	Information Economy Meta-Language
IoT	Internet of Things
IP	Internet Protocol
ITS	Intelligent Transportation System
JBRJ	Jardim Botânico do Rio de Janeiro
LMS	Learning Management System
LRD	Long Range Dependence
mLMS	Mobile Learning Management System
NER	Name Entity Recognition
NFC	Near Field Communication

NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OWL	Web Ontology Language
PACF	Partial Autocorrelation Function
POIs	Points of Interest
POS Tags	Part-of-speech tags
QALD-4	Question Answering over Linked Data - version 4
QAS	Question Answering System
RBM	Restricted Boltzmann Machines
RDF	Resource Description Framework
RFID	Radio-Frequency Identification
SLR	Semantic Role Labeling
SNR	Signal-to-Noise Ratio
SOM	Self-Organising Maps
SPARQL	Sparql Protocol and RDF Query
TARFIMA	Truncated Auto Regressive Fractionally Integrated and Moving Average
TF-IDF	Term Frequency - Inverse Document Frequency
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
USL	Uniform Semantic Locator
W3C	World Wide Web Consortium
WoT	Web of Things
WSN	Wireless Sensor Networks

Contents

Introduction **14**
0.1 Motivation **14**
0.2 Scope **15**
0.3 Research Questions **17**
0.4 Objectives **18**
0.4.1 The general objective 18
0.4.2 Specific objectives 18
0.4.3 Justification 19
0.4.4 Contributions to the field 20
0.5 Document Structure **21**

1 BACKGROUND AND STATE-OF-THE-ART **22**
1.1 Chapter overview **22**
1.1.1 An individual-centered approach 23
1.2 Individual data **23**
1.2.1 Individual data found on Social Networks 26
1.2.1.1 Characteristics of Twitter 26
1.2.2 Self-Organizing Map - SOM 27
1.2.3 Sentiment Analysis over Individual Data 28
1.3 Sensor Data & Internet of Things **29**
1.3.1 Sensor data and noise 29
1.3.1.1 Predictions based on Sensor Data 30
1.3.2 A Semantic Meaning to Sensor Data 30
1.3.3 Virtual Sensors to aggregate sensor data 31
1.4 Web Data **32**
1.4.1 Question Answering System - QA System 33
1.4.2 A semantic meaning to Web Data 35
1.5 Correlated works **36**

2 CHARACTERISTICS AND ISSUES OF HETEROGENEOUS DATA SOURCES **40**
2.1 Introduction **40**
2.2 Characteristics of data created by Individuals **40**
2.2.1 Denoising Individual data 44
2.3 Characteristics of Sensor data **44**
2.3.1 Denoising sensor data using Discrete Wavelets transform 46

2.4 Characteristics of Web data . . . 46

3 SOFTWARE DEVELOPMENT LIFECYCLE . . . 49
3.1 The overall description . . . 49
3.2 Individual data at the heart of the Design . . . 49
3.2.1 Module 1 - Individual Data - General Requirements Specification . . . 50
3.2.2 Module 2 - Sensor Data - General Requirements Specification . . . 52
3.2.3 Module 3 - Web Data - General Requirements Specification . . . 56
3.2.4 Module 4 - Process of Integration . . . 57
3.2.5 Module 5 - Applications: Mobile Applications and Dashboards . . . 58

4 DEVELOPMENT & VERIFICATION . . . 60
4.1 Individual Data - Twitter and Self-Organizing Maps . . . 61
4.2 Sensor Data: Denoising sensor signal using Discrete Wavelet Transform . . . 63
4.2.1 Methodology of the Experiment . . . 64
4.2.2 Application . . . 69
4.2.3 Final discussion . . . 72
4.2.4 Predictions based on exploratory data analysis . . . 73
4.3 Web Data: Question Answering (QA) System . . . 76

5 SMART SCENARIOS . . . 79
5.1 Botanical Garden of Rio de Janeiro digitally augmented . . . 79
5.1.1 Possible solutions . . . 80
5.1.2 Virtual Environment . . . 81
5.1.3 Physical Environment . . . 83
5.1.4 Citizen data: Understanding data in learning and e-learning contexts. . . . 84
5.1.5 Possible benefits of the system . . . 86
5.1.6 Research & System Validation . . . 86
5.2 Using wearable devices to measure local happiness. . . . 88
5.2.1 The World Happiness Report launched by United Nations . . . 88
5.2.2 The suggestion of an application . . . 88
5.2.3 Data coming from Social Networks . . . 89
5.2.4 Data coming from wearable devices . . . 89
5.2.5 Data coming from web sources . . . 89
5.2.6 Data analysis and technical view . . . 90
5.2.7 Steps to connect the application . . . 90
5.2.8 Benefits to the user . . . 90
5.2.9 Wearables devices that will be tested in this scenario . . . 91

5.3 Data fusion in intelligent transportation systems: using the citizen as a sensor . . . 92
5.3.1 Requirements identification . . . 92
5.3.2 Main characteristics of the proposed tool . . . 93
5.3.3 Description of the platform proposed . . . 94
5.3.4 Limitations and future works . . . 97

6 CONCLUSION . . . 98
6.0.5 Contributions . . . 99
6.0.6 Future Work . . . 100
6.0.7 Final Considerations . . . 100

References . . . 101

APPENDIX 113

APPENDIX A – PUBLISHED PAPERS . . . 114

APPENDIX B – POSTER PRESENTATIONS . . . 115

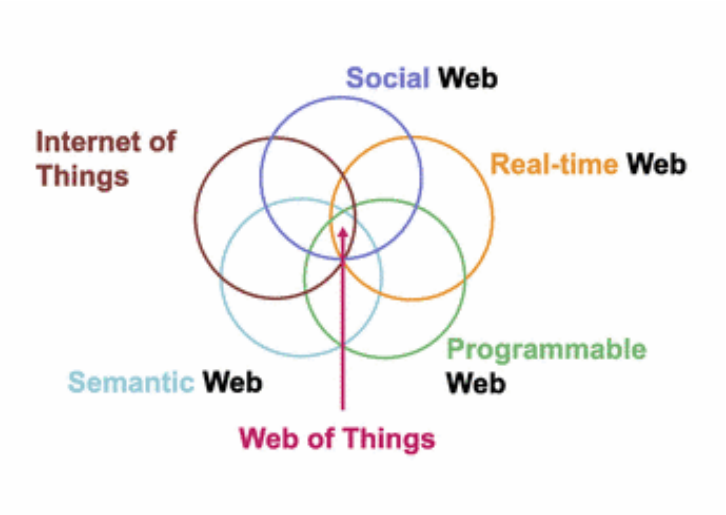
Introduction

The rapid advancement of hardware and software as well as the concepts of ubiquitous computing technologies are providing more personalized and intelligent services. This research aims to contribute to the development and the optimization of these services in order to enhance the experience of the user. This chapter provides an overview of the scope and goals of this study.

0.1 Motivation

As shown in Figure 1, the Web of Things (WoT), which encompasses several concepts like Internet of Things (IoT), Social Web, near Real Time Web, Programmable Web and Semantic Web, allows data to be published from several sources.

Figure 1 – WoT encompasses data from several sources



Source: Guinard (2010)

This implies that the volume of information on the Internet has a huge potential to grow exponentially enabling the access to a large amount of data from various physical, digital and social sources. Cisco IBSG forecasts that there will be 25 billion gadgets connected to the Web by 2015 and 50 billion by 2020 (EVANS, 2011). In fact, not only are individuals producing new data to the Web 2.0, but also objects known as "things" like smartphones, sensor nodes, tablets, GPS, smart glasses, watches, bracelets, rings, hair clips and others. These devices undergo constant technological innovation and subsequent

reduction in price. Therefore, individuals are acquiring and employing them as an aide to their personal activities.

In general, people describe facts, express opinions and sentiments in social networks, blogs, microblogs, bookmarks, news and events, images, video, audio and comments about products. Glorot, Bordes e Bengio (2011) performed their research using the data set of Amazon, which has more than 340,000 reviews, covering 25 domains, for 22 distinct product types. In general, people usually send and receive data from the Web, creating new data from simple sensor records to complex computer logs.

In addition, the Internet of Things have been widely disseminated and investigated, not only by academics but also by various segments of industry and commerce (SHULL, 2013), (EVANS, 2011), (BIMSCHAS et al., 2011). It is likely that the use of sensors will be massive and they will be expected to understand, monitor and actively respond to major events. When all sensors are in place and fully operational they will send a large amount of data to be analysed and understood in near real-time. These data will flow at high speed, in large volume, varied in type and value of information.

The scenarios described above shows that, data and information can be generated at any moment, in anyplace, by anyone or anything. The varied number of data sources will produce different data types, each one describing specific details from which it will be possibly necessary to extract information. As data proliferate and volumes grow, the extraction of relevant information can become very difficult, because a significant number of data of little importance emerges alongside the main information or data can be polluted with noise.

This work investigates three heterogeneous data sources: individual data, sensor data streams (also sensor signals) and web data in order to identify their specific characteristics and issues, particularly the problem of signal noise. This study also investigates Discrete Wavelet Transform (DWT) to reduce noise. Results here show significant improvements in the sensor signal when using DWT. The principal conclusion of our analysis is that, by knowing the main issues of data, it is possible to track problems and reduce data errors. Another practical conclusion is that DWT is a powerful tool for reducing signal noise.

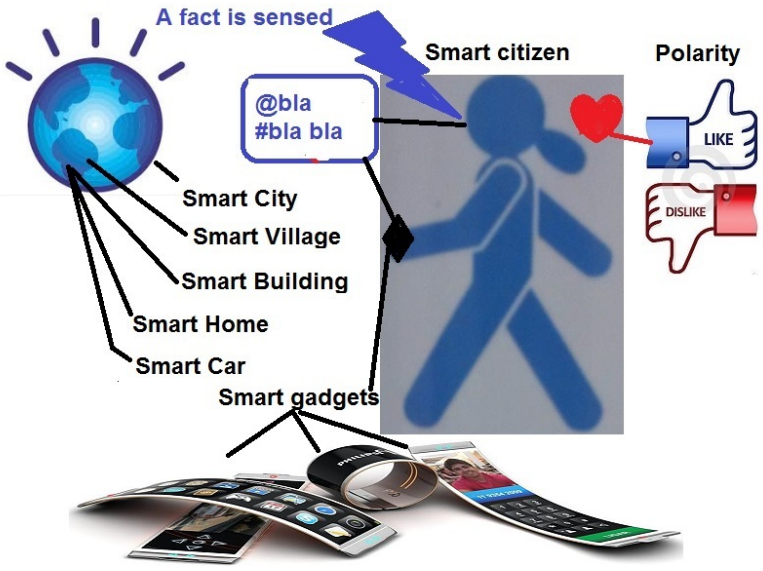
0.2 Scope

In recent years, new concepts and paradigms have been emerging due to expansion of technological innovation. Examples are smart cities (ROSCIA; LONGO; LAZAROIU, 2013), smart homes (KAMILARIS; PITSILLIDES, 2013), smart cars (MAMMERI et al., 2013), smart devices (SUAREZ-TANGIL et al., 2014), smart citizens (LIYANAGE; MARASINGHE, 2013), etc. All these novelties are connected to the Web receiving/sending information and generating a huge amount of data. Figure 2 shows that, when a fact is

sensed an individual can obtain information from diverse sources immediately, anywhere and anytime. This implies that he/she can make decisions and share opinions based on that information. If the system understand these actions and feelings it can adjust itself to fit into the new context.

It follows that, this smart universe (described in Figure 3), which includes the individual (his/her sentiments) and the varied types of "smart data" generated around them are investigated here. If a system can understand what type of data is being sensed and merges them with different data sources, it can provide a wide perspective about an occurrence and improve the experience of the user.

Figure 2 – The smart universe around the smart citizen.



Source: The author.

Figure 3 shows that, to obtain a clear understanding of a fact, it is necessary to collect and interpret data from varied sources simultaneously. Autonomous data fusion and its correct interpretation might allow a clear understanding of a occurrence, provide customised services and enhance the experience of the user.

However, traditional systems are not ideally suited to respond in accordance with the event. While Jones (2014) states that *"systems are still in their infancy"*, Chen e Lin (2014) believe that *"it is not an ordinary task"*. In general, it is felt that further research is needed in aggregation of heterogeneous data types, in order to provide to the systems with an understanding about the user, his environment and ultimately, deliver personalised information. Consequently, this study aims to go one step further in terms of autonomous discovery of information. It highlights the importance of data and investigates several techniques to extract intelligible information from heterogeneous sources of data in order

Figure 3 – Systems must be improved in order to understand not only the environment, but also the user and his sentiments.

Source: The author.

to provide useful information and customised applications. According to Sabosik (2013) *"How aggregated data will be analysed will be a topic of discussion in the future."*

On the other hand, this study excludes the analysis of the financial and psychological impact caused by the facts discovered. It also does not solve real problems found during the process. It only provides information and system actuation according to specific occurrences promoting a wide view to the user. Moreover, it does not investigate hardware and communication infrastructures, nor security, confidentiality and privacy. Finally, it does not suggests software or hardware. It encompasses techniques that can enable an understanding of a fact and sentiments involved in order to adjust the systems to work according to that occurrence.

0.3 Research Questions

The scenario described in this section suggests that future research should focus on developing combined techniques to analyse raw data, discover meaning, and deliver information to the user (individual) on time. Therefore, the research questions in this study are:

1. How can public data (data created by individuals) be collected, analysed and clustered in order to group similar profiles?

2. How can sensor data be denoised and reduced enabling a more trustworth extraction of data?

3. How can users access information in the web easily and quickly, posing questions in Natural Language?

4. What is the best scheme to integrate heterogeneous data sources?

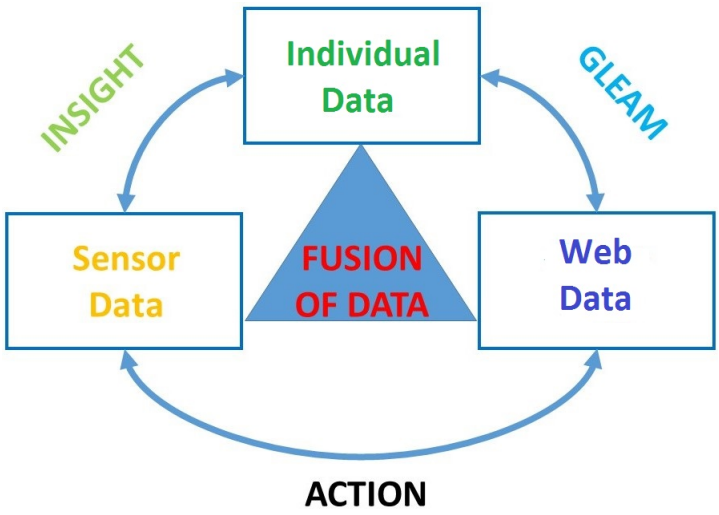
0.4 Objectives

The objective of this thesis is twofold. It is divided in two subsections: general objective and specific objectives. Next, it is presented the justification and contributions to the field.

0.4.1 The general objective

The main objective is to investigate three different data sources (individual data, sensor data, web data) and techniques to deal with them. This work includes research into techniques that process, interpret and reduce these data. The proposed platform [1] delivers personalised information (which constantly arises) to applications and services. Figure 4 shows the main idea embedded in this objective.

Figure 4 – Data fusion can provide better information, resulting in wise ideas and better decisions.



Source: The author.

0.4.2 Specific objectives

In this study, a set of techniques for dealing with heterogeneous data types is presented. Together, these techniques will compose a platform that aggregate and deliver

[1] From now on the word Platform will replace the word Framework in accordance with recommendation obtained during the Viva Presentation.

personalised information. More specifically, the objectives to be addressed in this study are:

1. An investigation into three different data sources (individual data, sensor data, web data) is proposed here. The following hypothesis is made: when fused, these three data sources can provide a wide perspective of an event and personalised information.

2. Each of these data sources: individual data, sensor data, web data, contribute with heterogeneous data types. Because of this, it is necessary to investigate different methodologies and techniques to deal with these three sources. Afterwards, these techniques must be aggregated in a platform in order to work together (extracting and delivering information to Apps and dashboards).

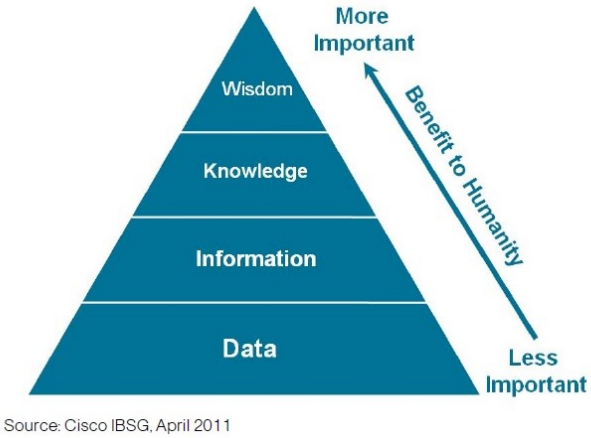
3. The countless number of sensors, simultaneously and automatically producing signals (that will be transformed into large amounts of data) means that tremendous effort is required in collection, processing, storage and retrieval of the data. Such huge volumes of data can be reduced if the underlying signal is denoised. Therefore, we propose an innovative technique to improve the signal of a sensor, enabling good results when employing a filtered sensor signal in applications.

0.4.3 Justification

Although many studies have been focused on the main issues regarding the data: i.e. volume, variety, velocity, veracity, variability, visualization and value (SHULL, 2013), (DEVLIN; ROGERS; MYERS, 2012), (EVANS, 2011), (BARNAGHI; SHETH; HENSON, 2013), very little research has been done on the aggregation of heterogeneous data types that consider, not only the environment, but also the user (individual). In addition, Rantanen e Sillberg (2014) state that *"Different data sources are often provided in various formats, thus it may be difficult to parse the same data from them all"*. Consequently, this work investigates the fusion of three data sources: data produced by individuals, sensor data and web data. When these three sources are aggregated new perspectives can arise, and it might be useful to the user. As showing in Figure 5, data is an essential component in the process of acquiring wisdom. In other words, taking decisions based on secure sources of information can be useful to the individual.

For all these reasons, the aim of this research is to enhance the experience of the individual by investigating new methods of extracting information from raw data, finding what real analytic and operational value can be mined from it. Here, techniques, algorithms, and new approaches are suggested in order to provide meaningful applications and automated decisions for near real-time process, better planning, forecasting and actuation. According to Sheth, Anantharam e Henson (2013) *"knowledge is generated by*

Figure 5 – Turning Data into Wisdom



Source: Evans (2011)

continuously observing human activities within physical-cyber-social worlds, and we can use this knowledge to improve human experience."

0.4.4 Contributions to the field

This thesis presents a novel process that has shifted its focus from IoT technology to the user (or Individual/smart citizen). This process takes into consideration data created and shared by Individuals, meaning that this source of data is the most important. Correlated data (i.e. data from sensors and from the web) must be aggregated to Individual data. The major benefit to the user is that the proposed platform delivers information in accordance with individual's perspective. In other words, it is the basis of individually and socially aware applications and services. Also, the research outlines several characteristics of three different data sources in order to show that there are numberless issues that needs to be tacked.

Besides the contributions of the process, that employs IoT but focus on the user, a scheme for the first phase of the IoT process is presented here, when sensors are sensing and transmitting data in signal format. This research presents several reasons why a signal has to be analysed, denoised and filtered. It shows that huge volumes of data can be reduced if the underlying signal has adequate spectral properties and good results can be obtained by employing a filtered sensor signal in applications as shown in Chapter 4. Finally, it highlights that if the noise does not compete with the desired signal, the extracted information can be more precise.

0.5 Document Structure

This initial chapter introduced the motivation and scope, defined the boundaries, research questions and their linked objectives. Chapter 1 presents the platform divided into four parts, the main issues associated with three data sources and correlated platforms. Chapter 2 highlights the main characteristics of each data source and a preliminary investigation of noise. Chapter 3 describes the Software Development lifecycle divided in modules. Chapter 4 presents the main findings of this thesis. Chapter 5 indicates three possible scenarios where the proposed platform might be useful. Finally, Chapter 5 discusses the main findings and concludes this thesis.

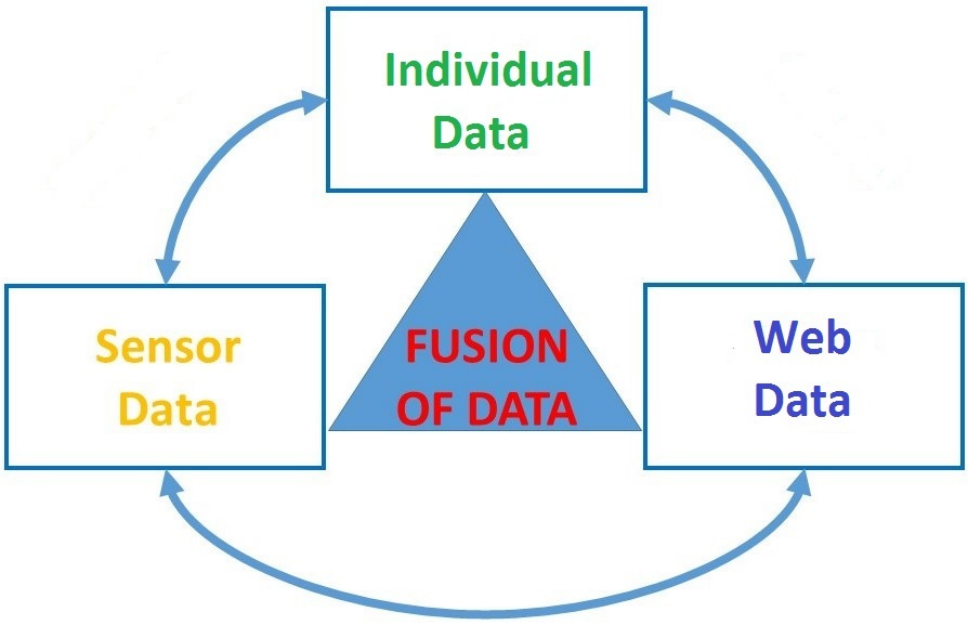
1 Background and State-of-the-Art

This chapter investigates the main issues related to the distinct areas of the platform, and techniques to deal with them that would produce the best outcomes.

1.1 Chapter overview

The fusion of three data sources is proposed in this research: individual data, sensor data and web data. All these three sources make use of the Internet, however each of them have aspects that are very distinct and diverse. Although they ultimately have the same destination (the Internet), it is necessary to evaluate these data separately and differently. To better explain the rationale for this structure, this chapter is divided into four sections and explores the state-of-the-art in the areas depicted in the Figure 6. The first section describes the core of Individual data. The proposed platform puts a big emphasis on the data posted by individuals, who are also called smart citizens. The second section deals with sensor data and Internet of Things. Section three describes briefly the basic of Web Data and Question Answering System. Finally, the last section provides an overview of correlated platforms.

Figure 6 – Scheme of organization.



Source: The author.

1.1.1 An individual-centered approach

It is notable the change in the societal organization. Marsal-Llacuna e Fabregat-Gesa (2016) proposed a citizen-centric urban planning arguing that citizens have changed considerably during the last decade. However, urban planning processes are the same since since 70s. The author considered local governance and city policies when measuring the "citizen-centeredness" of sustainability. In their research, the citizen-centric approach is determined by monitoring efficiency of cities in safe-guarding citizenship rights (MARSAL-LLACUNA, 2016).

User-centered approach has been widely accepted because the major attention of the project goes to the user. According to Knoeri, Steinberger e Roelich (2016), end-users are the consumers of the end-use services, which are used to satisfy their needs and wants in accordance with their profile (i.e. lifestyle and human and financial capital). The platform proposed in this research take into consideration the requirements and needs of the Individual, giving primary attention to their data when integrating related data. The platform proposed in this research take into consideration the requirements and needs of the Individual, giving primary attention to their data when integrating related data. It means that the fusion of data proposed here will produce an Individual overview.

Richter e Flckiger (2014) mention several fields using a user-centered approach, among them are: Human Computer Interaction (HCI), Human Factors, Interaction Design, Usability Engineering, User-centred Design (UCD), User Experience (UX). These fields of research have in common the purpose of systematically developing and improving products for the individuals who will use them.

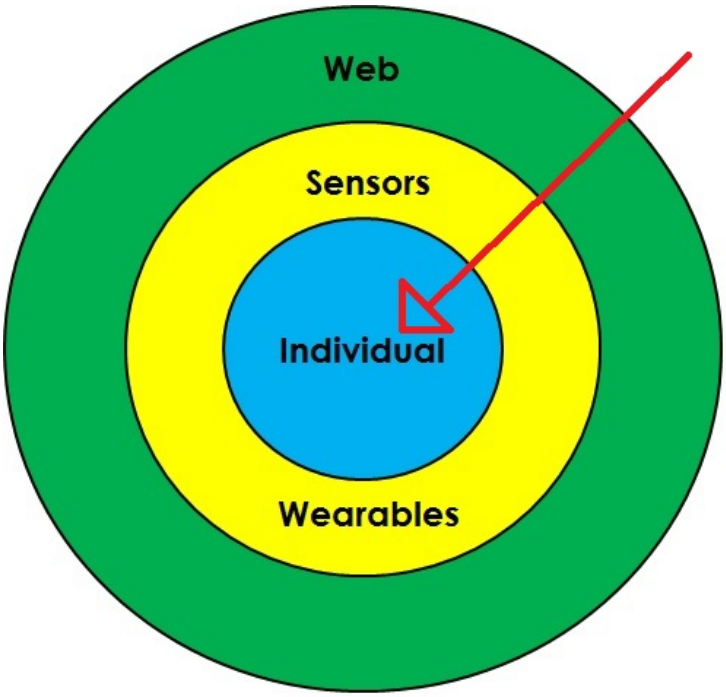
According to Fauquex et al. (2015) *"IoT is currently at a turning point where the hardware and technology are mature and the focus can be put on creating good user experiences."* For all these reasons, the platform proposed highlights the importance of Individual's data delivering information in accordance with their perspective.

1.2 Individual data

The world is becoming smarter. Gobbi e Spina (2013) state that smart means customised solutions, adapted to the situation and flexibility based on a dynamically connected society. According to them, smart cities are not only interconnecting with each other, but also connecting us with the environment by means of technology. Liyanage e Marasinghe (2013) indicate that a Smart Citizen is someone active, having good principles and creating smart schemes for all tasks in the best way.

Usually, a Smart Citizen is using one or more devices to collect and send information. Due to the vast amount of gadgets available today, citizens are being considered active

Figure 7 – The individual perspective is the target. Correlated data must be aggregated taking into consideration firstly individual data.



Source: The author.

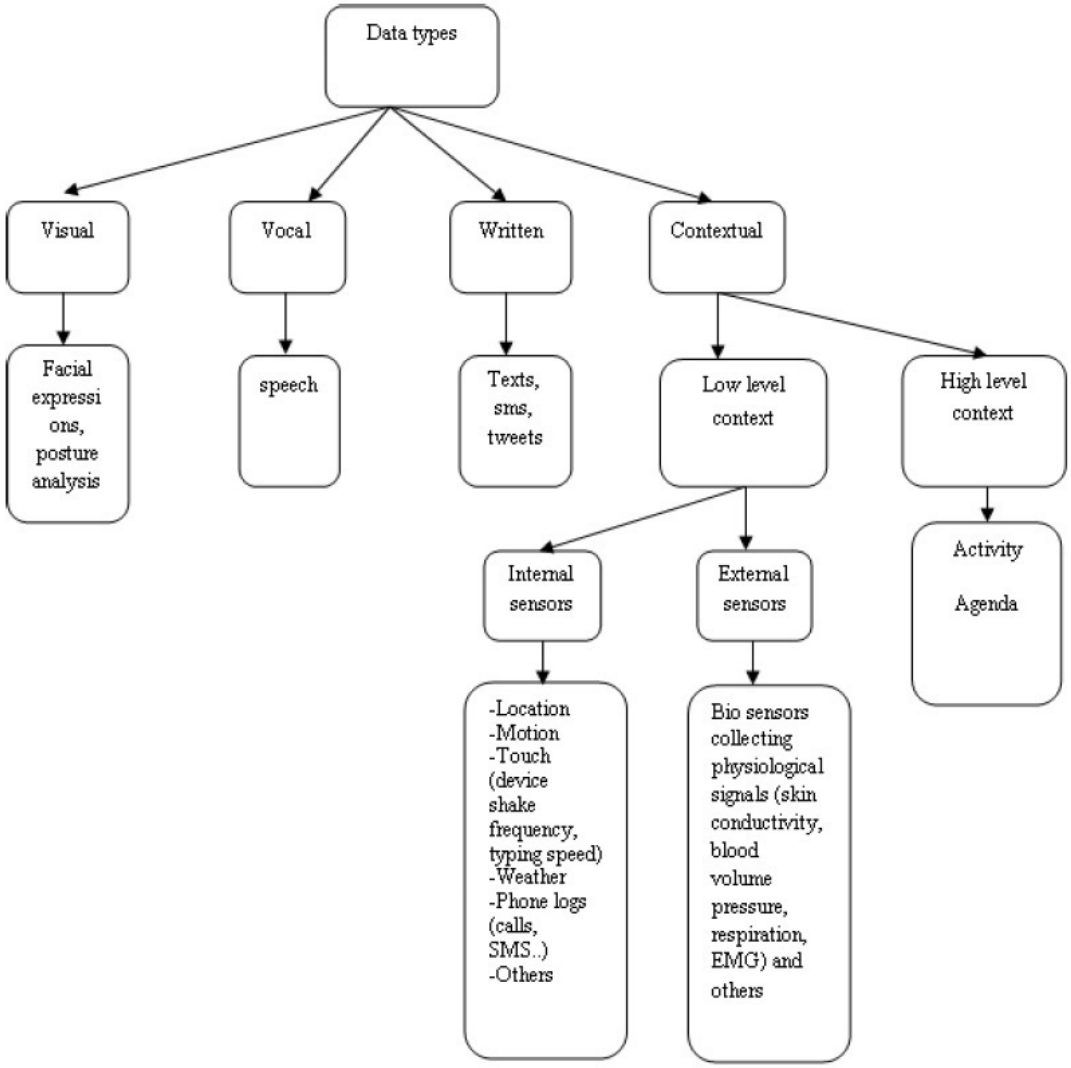
and passive sensors increasing the quantity of information available on the web. Thus, a paradigm emerges: citizens as sensor, where citizens are voluntarily acting as a type of sensing device transforming facts in digital information and disseminating it in order to report or alert.

Boulos e Al-Shorbaji (2014) describe people in the loop proceeding as citizen sensors by collecting and sharing data using their own gadgets, and in doing so, they consume and also produce regional information. *Prosumer* is a term used to indicate this profile: producer and consumer simultaneously (GOBBI; SPINA, 2013). Therefore, any device-holder has the potential to become a lively proactive and perceptive sensor. Under these circumstances, a new term is being disseminated "Mobile Crowd-Sensing" where individuals with sensing and computing devices collectively share data and extract information to measure and map phenomena of common interest (GANTI; YE; LEI, 2011). For all these reasons, Smart Citizen, Citizens as Sensor, Mobile Crowd Sensing - MCS, are becoming not only popular expressions, but also an area of investigation.

Sprake e Rogers (2014) state that citizens use this mobile interaction to make judgements on what they are experiencing and learning by collecting, sharing, interpreting, refining, and examining what they sense. Recently, mobile devices have been offering integrated tools for harvesting, promulgating and subscribing to the web to aid networks

of citizens sensors producing data about the local environment.

Figure 8 – Data types noted by Mousannif & Khalil



Source: Mousannif e Khalil (2014)

On the one hand, citizens describe facts, share knowledge, express opinions and sentiments in social networks, blogs, microblogs, they write comments on magazines, stores online, etc. On the other hand, simultaneously they are producing diverse types of data. According to Mousannif e Khalil (2014) these data could be visual, vocal, written or contextual. Figure 8 maps the most common data type produced by citizens.

For all these reasons, an Individual-centered approach should be take into consideration. The concept of Smart Cities 3.0, adopted by cities like Amsterdam, Copenhagen, Helsinki and others, give to the citizen extensive attention using citizen co-creation models, enabling collaborative communities, allowing efficient use of technology to help the development of the city.

1.2.1 Individual data found on Social Networks

Online Social Networks (OSNs) are now part of the life of people from all over the world (YU; LIU, 2011). According to Buccafurri et al. (2015), OSNs are the most popular communication media on the Internet. They help people keep their contacts online, interact with other users, spread ideas and news, share events and products, etc. Consequently, the volume of generated data (created by users) is exponentially growing. There are several of them: Facebook, Twitter, Instagran, LinkedIn, Google+, Flickr, Pinterest, etc. Each one has special characteristics and roles, captivating people in different stages of life. Consequently, social networks are attracting great interest from researchers of several areas.

In this research, we are interested in important aspects of emotion, mood and preferences. For example. Roshanaei e Mishra (2015) mention that current experience and feeling of the users can be found in emoticons and mood updated by users. Hennig et al. (2014) provide an overview about the emerging trend of topics; closest to this work is Saito, Tomioka e Yamanishi (2014) who are studying early detection of persistent topics. On the other hand, 120,000 fraudulent accounts were investigated by Thomas et al. (2013). For the most part researchers are using Twitter to perform different type of analysis. This preference has several reasons as demonstrated in the next subsection.

1.2.1.1 Characteristics of Twitter

Many researchers begin their analysis using Twitter, there are several reasons to make this choice:

- Tweets are restricted to a maximum of 140 characters while Facebook entries can be voluminous (KAISLER et al., 2014);
- It has a quasi-realtime nature (BEYKIKHOSHK et al., 2014);
- Its content is mostly public and very concise (CHEN; VORVOREANU; MADHAVAN, 2014);
- Twitter is a popular social network (CHEN; VORVOREANU; MADHAVAN, 2014)(SAITO; TOMIOKA; YAMANISHI, 2014);
- It has a button called retweet allowing users to share any tweet posted by other users (SAITO; TOMIOKA; YAMANISHI, 2014);
- Twitter is being used for different purposes such as to connect people with the same interests, communicate with celebrities, product marketing, spread Internet-based phenomena (memes), political campaigning, citizen journalism, market research, etc. (CHEONG; LEE, 2010);

- Tweets are free of cost, are visualized world-wide and posted in real time (JI et al., 2015);
- Twitter has an intelligent classification scheme using hashtags (KAISLER et al., 2014);
- There are more than 70 different languages in this Social Network.(MOCANU et al., 2013)

The list above has been taken into consideration and the first prototype, developed during this research work, uses data harvested from Twitter during the analysis.

1.2.2 Self-Organizing Map - SOM

After collecting data from Twitter, an exploratory data analysis using Self-Organizing Map (SOM) is applied to the data set. SOM is an unsupervised learning technique, mainly used in the exploration of large text collection. SOM is applied not only with analytical purposes, but also as a visualization tool. This method shows a kind of similarity diagram of the models, in order to obtain an insight into the topographic relationships of data, particularly high-dimensional data items (KOHONEN, 1998).

In this research context, SOM is useful because Social Network are organized by several thousands of nodes (actors) and edges (ties) that connect nodes in various relationships (AKAY; DRAGOMIR; ERLANDSSON, 2015). SOM is capable of clustering these nodes according to some similarity (correlations among profiles found in Twitter) or distance measurement. In summary, SOM executes the clustering process based on similarities found on Twitter's profiles. The clusters, formed during the training stage, can be analysed in order to determine what has been discovered.

Prieto et al. (2007) applied SOM when analysing blogs hosted by Blogalia. The authors aimed to discover communities, i.e. blogs with similar link characteristics. According to them, SOM was useful to identify and visualise emerging communities. In the same way, Cheong e Lee (2010) applied SOM, in conjunction with other technology, to discover user sentiments, habits and demography when users are collaborating with topics of discussion on Twitter. The authors found hidden patterns for topics of interest when applying a methodology that was more analytical and objective.

Kohonen (2013) highlights that one of the main advantages of the SOM algorithm is that one can compute really large mappings in short time, using a personal computer. For this reason, SOM is the chosen algorithm to map communities in Twitter during the prototype phase.

1.2.3 Sentiment Analysis over Individual Data

Chambers Dictionary define sentiment as an opinion, attitude or judgement; a thought or feeling expressed in words and influenced by emotion. However, to better understand this important topic, it is necessary to review the clinical terminology. In his paper, Levine (2007) provided two interesting definitions for emotion:

"Emotion tends to be used for what a person is feeling at a given moment. Joy, sadness, anger, fear, disgust, and surprise are often considered the six most basic emotions, and other well-known human emotions (e.g., pride, shame, regret, elation, etc.) are often treated as elaborations or specializations of these six when occurring in complex social situations."

"Emotion is that which leads one's condition to become so transformed that judgement is affected, and often accompanied by pleasure or pain." Aristotle's statement *(in Rhetoric)* quoted by Levine (2007).

There are several advantages in performing sentiment analysis of Internet users. First, people express their opinions and judgement about commodities and services that they are aware of. Thus, it is possible to estimate the efficacy of the product or service (ROUTRAY; SWAIN; MISHRA, 2013). Bollegala, Weir e Carroll (2013) highlight that reviews of a large amount of products are available on the web such as, readers opinions (amazon.com), guests comments (tripadvisor.com), movies rating (imdb.com), users of cars (caranddriver.com), opinions about food (yelp.com), etc. Second, public sentiment and emotion regarding topics and facts and the way this is delivered and expanded, is of particular interest in many contexts because it is fast and cheap (COLBAUGH; GLASS, 2013). Third, a politician can modify his/her attitude according to public sentiment (TAN et al., 2014). Finally, the polarity of the sentiments disseminated by students registered in a course can create a useful form of feedback to the tutor.

Many other areas are being investigated (SALOUN; HRUZIK; ZELINKA, 2013), (MAKAZHANOV; RAFIEI, 2013), (YANAGIMOTO; SHIMADA; YOSHIMURA, 2013), (ORTIGOSA; MARTIN; CARRO, 2014). Bollen, Mao e Pepe (2011) demonstrate that mood analyses offers a reliable framework to model a tendency in collective emotions generating a valuable predictive insight of social and economic issues.

1.3 Sensor Data & Internet of Things

Due to the phenomenon called Internet of Things, not only ordinary devices like computers, tablets and smartphones are generating data, but also cars, houses, camcorders, sensors, devices for navigation and location, etc. These elements are being called things. According to Kenda et al. (2013) the things are devices that can be digitally labelled using technologies like IP v4 or v6, NFC, RFID, EPC, etc.

Things digitally labelled are progressively being adopted to monitor, collect and send data about many events occurring in real time, for example, from traffic movement to power utilization, from ocean wind to temperature, from presence to heart beats. It is likely that there will be considerable benefits when all these technologies are established and in full use. For example, while Min (2013) indicates that sensors can reduce the costs of distribution and logistics and provide effective product tracing, Asin (2012) highlighted far more benefits by cataloguing 54 domains under twelve categories.

These things and sensors (including citizens as sensor) are generating a large variety of data. Consequently, the quantity of data is growing exponentially generating a new concept called Big Data. Boyle, Yates e Yeatman (2013) state that IBM categorises the dimensions of Big Data into Volume, Velocity, Variety and Veracity. These four concepts can be applied to the sensor data sphere. In addition to Volume, sensor data may change continuously and have to be processed in near real-time – as a Velocity issue. Sensor data are generated from a Variety of heterogeneous systems leading to a variety problem. Data accessibility beyond silos of heterogeneous systems can lead to a Veracity problem" (BOYLE; YATES; YEATMAN, 2013). These characteristics indicate that it is a big challenge to extract knowledge from sensor data.

1.3.1 Sensor data and noise

Sensor data faces many challenges, they are listed in chapter 2, section 2.3. In this research, a scheme for the first phase of the IoT process is presented, when sensors are sensing and transmitting data in signal format. Recently, many authors recommended different methods to remove the noise of the sensor signal (ZHANG et al., 2012) (MONTE et al., 2013) (FACCHINETTI; SPARACINO; COBELLI, 2010). Wang, He e Chen (2009) applied three different techniques, Wavelet Shrinkage Threshold (WST), General Matching Pursuit (GMP) and Genetic Matching Pursuit (GAMP), to remove the noise of some typical signals from stationary and non-stationary engines. The authors were specially concerned with accuracy and computation speed. However, in the IoT context, the main concern is the volume of the data and the challenges in processing it. Because of this, an investigation into the sensor signal using Discrete Wavelet Transform (DWT) is proposed. The technique presented in this thesis may drastically reduce the volume of the data if the

signal has adequate spectral characteristics. In this case, it is possible to reconstruct a new compact representation of the signal. This technique is also useful to remove the noise.

Noise is defined as a signal component that does not have a strong correlation with any vector of the dictionary (YE; DOBSON; MCKEEVER, 2012). In other words, noise refers to any unwanted change that modifies the values of the signal (BERKNER; WELLS, 1998). This feature can lead to several problems, because noise will require extra time and resources for processing useless unwanted data. Percival e Walden (2000) demonstrate that the DWT can be a useful way of representing (re-expressing) the signal in a manner that will help the problem of signal estimation. According to Debnath (2001) thresholding the noisy wavelets coefficients, using the right thresholding scheme, removes most of the noise and preserves the large coefficients. Thus, we use DWT as a tool to denoise the sensor signal.

1.3.1.1 Predictions based on Sensor Data

Wireless Sensors Networks (WSNs) allow the observations of various phenomena by providing constant measurements (SANTINI; RöMER, 2006). It means that the use of communication will be massive, which affects the lifespan of sensors. To tackle this problem, predictors were suggested to the field of WSNs.

Benzing et al. (2010) predicted values instead of actual readings are useful to reduce the amount of data sent to the WSN gateway at the expense of slightly less accurate values.

Lazaridis e Mehrotra (2003) emphasize that sensor can compress time series instead of sending them in raw form and these series can be stored in Database systems. In this research a very basic investigation, using Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF), Variance and Periodogram is adopted. Although these techniques are very basic, they are essential for the analysis. That was a very interesting experiment and more research must be done in the future.

1.3.2 A Semantic Meaning to Sensor Data

According to Compton et al. (2009), as huge amounts of sensor data are ready for use on the web, semantic sensors statements and sensor data create ways to make such data visible, reachable, and queryable. In addition, Semantic annotation of sensor data presents a means of correlating the data to the semantic description. Moreover, using the principles of Linked Data simplifies the integration of stream data for an increasing number of Linked Data collections (PHUOC; PARREIRA; HAUSWIRTH, 2010), (SEQUEDA; CORCHO, 2009), (SHETH; HENSON; SAHOO, 2008).

Many works use the traditional approach of Semantic Web and Linked Data to sim-

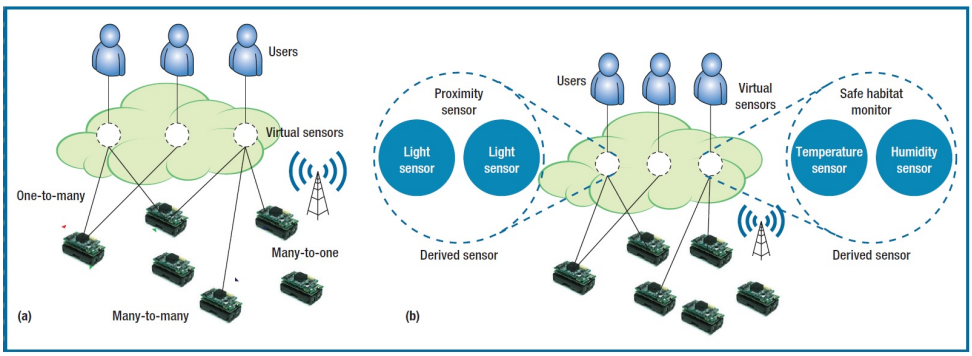
plify IoT applications development (BIMSCHAS et al., 2011). For example, Bhattacharya, Meka e Singh (2007) developed a distributed and hierarchical index structure for sensor networks based on statistical models. This Model-based Index STructure (MIST) can answer semantic queries based on two approaches: first transforming the readings into symbolic models locally at the sensors through semantic interpretation, second performing an in-network indexing and aggregating semantic models to capture the global patterns. This scenario reinforces the importance and difficulties of dealing with reliable and updated data.

Applying semantic technologies to IoT facilitates effective data access and integration, resource discovery, semantic reasoning, and knowledge extraction (BARNAGHI et al., 2012). Even though it is a promising idea, it still presents several difficulties and challenges. The Semantic Web uses ontologies to describe the semantics of many types of data. A key challenge in creating Semantic Web in the IoT context is the discovery of some specific data in large-scale IoT frameworks.

1.3.3 Virtual Sensors to aggregate sensor data

Virtual sensors can be a good alternative to aggregate sensor data. Kabadayi, Pridgen e Julien (2006) states that a virtual sensor is the opposite of physical sensor. It is a software sensor that emulates the same characteristics of a physical sensor combining sensed data from diverse sensors. However, it presents only customized data to the users. The virtual sensor includes metadata about the current users and the physical sensors. Moreover, when the user has complex queries the virtual sensor can present additional processing code in order to provide the answer. Madria, Kumar e Dalvi (2014) have started a network of virtual sensors in four distinct formats: many-to-one, one-to-many, many-to-many, and derived levels. Figure 9 shows the architecture of the mentioned schemas.

Figure 9 – An architecture using physical and virtual sensors



Source: Madria, Kumar e Dalvi (2014)

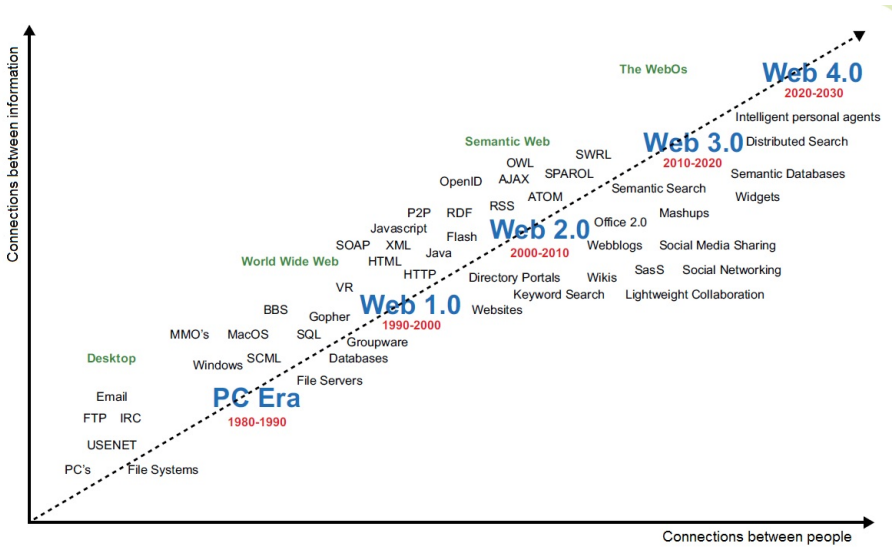
In this work, virtual sensor will use data from physical sensor and experimentally the concept of citizen as a sensor, meaning that the data around the citizen will be collected and these two data sources will be combined before the extraction information.

1.4 Web Data

The evolution of the Web is remarkable. In thirty years of existence, the web has transformed the shape of services and the way people communicate. Several technologies have been created and developed in order to provide better connection and information among people. Figure 10 shows this evolution: from the PC Era to Web 4.0. Around the central line, a group of techniques is arranged in ascending order.

As shown in Figure 10, several tools and technologies have been developed. These engines work together, are integrated and interact with users, delivering information as accurately as possible. The connections between technologies improves the connections between people by providing information anytime and anywhere.

Figure 10 – Chronology of Internet and Web evolution.



Source: Endres-Niggemeyer (2013)

Shull (2013) reported the following facts:

"We create 2.5 quintillion [1018] bytes of data every day, with 90% of the data in the world created in the last two years alone... Every hour, Wal-Mart handles 1 million transactions, feeding a database of 2.5 petabytes [1015 bytes], which is almost 170 times the data in the Library of Congress. The entire collected junk delivered by the U.S. Postal Service in one year is equal to 5 petabytes, while Google processes that amount of data in

just one hour. The total amount of information in existence is estimated at a little over ... a zettabyte."

Laender e Silva (2008) emphasize that the expansion of the Web has evolved into an immense repository of significant data for a variety of domains. Consequently, at the end of the line is placed "Intelligent personal agents" which is necessary to satisfy personal interests in a timely manner (KAWTRAKUL et al., 2014). As noted by Riedel, Memon e Memon (2014), these *"Intelligent agents"* can only work appropriately when expertise of diverse fields, such as machine learning, data mining, data management and parallel processing are adequately combined. Under these circumstances, the term Data Analytics has emerged. This term is employed when is necessary to extract knowledge from huge amounts of data internally or externally available.

Sabosik (2013) believed that customer data analytics would be a principal centre of attraction in the business world for 2014. In fact, companies have been developing virtual tools to collect information about their customers. In the same way, governmental institutions have been collecting information about the environments and citizens and providing cyber novelties. For example, in May 2014, Rantanen e Sillberg (2014) reported the following information:

"Several nations in Europe have released formerly closed data sources collected by government agencies and municipalities. These open data initiatives have enabled new and innovative services. "

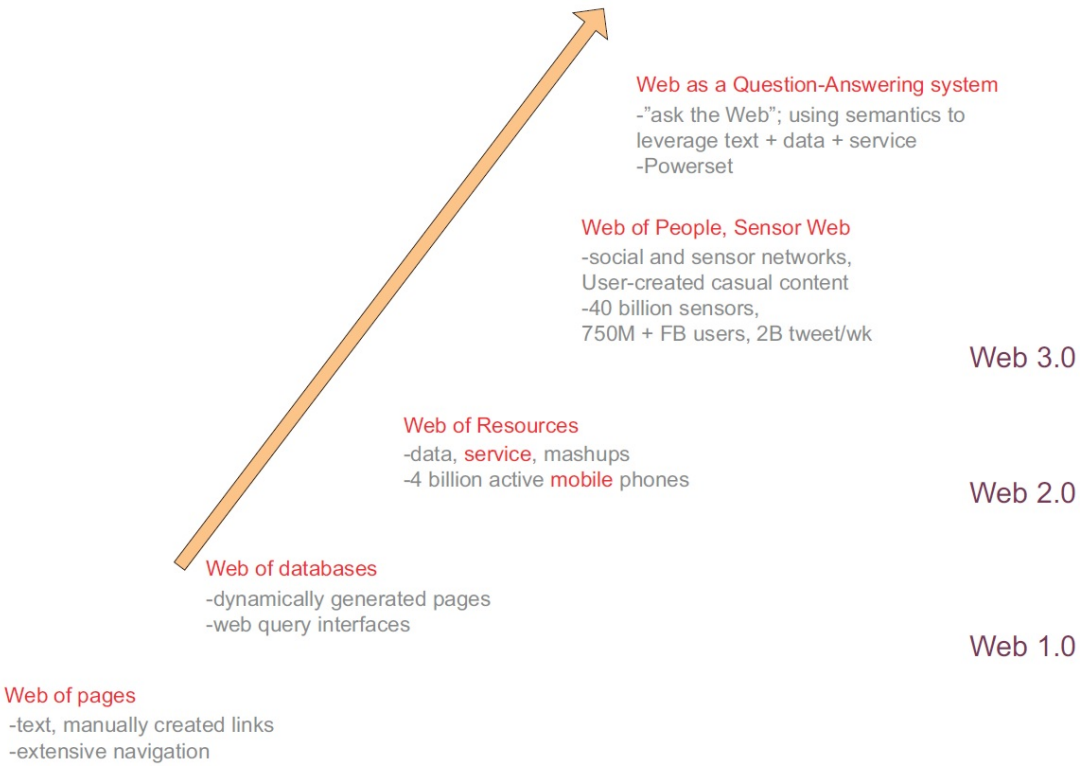
Cruz-Benito, Garcia-Penalvo e Theron (2014) state that it is necessary to design new schemes to collect and treat the data in order to facilitate data analyses and data mining. The author suggests the use of RDF, from W3C, to enrich data semantically as much as possible in order to allow a simple and structured understanding of the information it contains. The next subsection provides more information about the Semantics methodologies published by W3C.

1.4.1 Question Answering System - QA System

Usually, individuals share knowledge and ask questions using Natural Language, for this reason many systems are being developed to answer questions in natural language. These systems are called Question Answering Systems (QAS or QA sytems). The main purpose of a QA system is to provide exact answers rather than just show links to documents. It is included in a multi disciplinary field because it involves information retrieval, natural language processing, artificial intelligence, software engineering, linguistic knowledge, linked data and semantic web.

According to Tomljanović, Pavlić e Katić (2014), investigations of QAS initiated in 1960s, became more intense in 1970s, but the interest weakned in the 1980s. Nowadays, QA systems are becoming popular and it is positioned at the top of the Web evolutionary line in Figure 11.

Figure 11 – QA systems at the top of Web evolution.



Source: Sheth e Thirunarayan (2012)

There are several QA systems available. According to Visser (2012), Watson is one of the most effective systems, being able to answer various questions posed as phrases. Apple Siri is a QA system capable of sending email and texts, making calls to contacts, get directions, create reminders, create calendar events, search the web, open Apps, set alarms, adjust settings, play music, look up movie times, make conversions, etc. Microsoft Cortana calls itself a personal assistant. This system performs the same functions of Siri and is able to look up flights and recognise different music. Google now can share shortcuts, retrieve queries from background information and is useful for travelling. Amazon Echo is a dedicated device that provides information such as weather, dates, movies, etc. It is able to perform tasks like to-do list, playlist and adjust settings. There are others similar QA systems such as S Voice, Blackberry Assistant and Wolfram Alpha.

Question Answering over Linked Data (QALD) aims to evaluate current weakness and strengths of QA systems based on natural-languages over semantic data sources. QALD-6 is the sixth in a series that assess the ability of the systems to cope with large

amounts of heterogeneous and structured data. Finally, QALD analyses the progress of the system over time.

Usually, individuals show great interest for QA systems, specially young people. As the popularity of such systems increases and the number of users is augmenting, it is important to include in the platform an answer driven approach, such as QA system. By accessing information easily and quickly it could be attractive to the public.

1.4.2 A semantic meaning to Web Data

Semantic signifies meaning. Meaning allows a more effective use of hidden data (HEBELER et al., 2009). The principal objective of the Semantic Web is to promote a distributed Web at the data layer instead of the appearance layer. Rather than having one webpage linking to another, one data item can point to another, utilizing universal references named URI (ALLEMANG; HENDLER, 2011). "The Semantic Web promises to alleviate users from complicated, slow, and time-consuming tasks" (BREITMAN; CASANOVA; TRUSZKOWSKI, 2006).

Nowadays, the web is called a Web of Documents, where a document is linked with several others. However, it is difficult for computers to understand the content of these documents. To solve this problem, the Semantic Web was designed to create standards to link data instead of documents. It is claimed that computers can understand and work better with data than documents.

Tim Bernes-Lee and W3C are working to establish a standard way to publish data on the Web. They created RDF in order to link data to each other, they call these connections Linked Data. SPARQL, is a standardised language to query Linked Data. W3C believes that Linked Data is a collection of best practices to publish and connect data on the web. According to them, there are four mandatory steps to publish semantic data (BIZER; HEATH; BERNERS-LEE, 2009) :

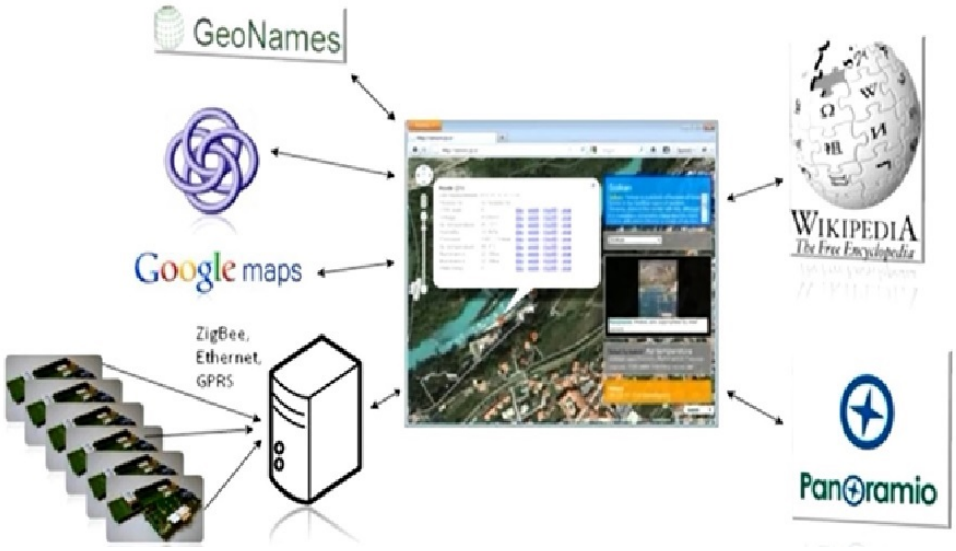
1. Use URIs as a name for things;
2. Use HTTP URIs so that people can look up those names;
3. Provide useful information using the standards (RDF, SPARQL) when someone looks up a URI,;
4. Include links to other URIs, so they can discover more things.

These four rules are proliferating around the world. Because various researchers are developing technologies to deal with semantic approaches (BIMSCHAS et al., 2011), semantic techniques will continue to evolve.

1.5 Correlated works

Kenda et al. (2013) describe the implementation of Videk, a prototype mashup for environmental perception and understanding. The user can observe the present condition of the system with the latest measurement or can visualize historical data from various sources, for diverse time periods in near real time. Basically, Videk encompasses sensor data and web data. This implies that, while some sensor nodes monitor streets and roads, other nodes measure conditions on rivers, temperature, humidity, etc. These sensors enable users to read realistic data distributed on the map and show the position of the sensor nodes (Videk is based on the Google Map API). There are six applications for location, cluster, photo, feature selection, event detection and contact. As shown in Figure 12 Videk aggregates physical data (sensor data) and web data, however, it does not use any citizen data.

Figure 12 – Videk enriches sensor data with web data, however it does not uses citizen data.

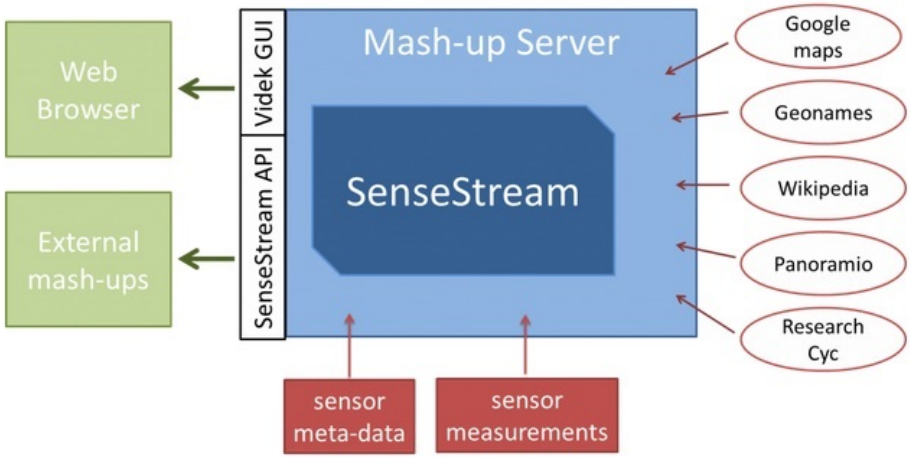


Source: Kenda et al. (2013)

Sensor data are enriched with different combinations of data, in order to permit the user to better interpret the measurements. In terms of system architecture, it uses four main components: external sources (Google Maps, Geonames, Wikipedia, Panoramio, Research Cyc), sensor data (meta-data, measurements), the user interface and the mashup server, as described in Figure 13. Videk is composed of modules. Consequently, more features and data sources can be added allowing a constant evolution.

SenseStream is the name of the storage and processing engine, which uses Apache as Mashup Server. On the other hand, GUI uses jQuery for data manipulation and Ajax

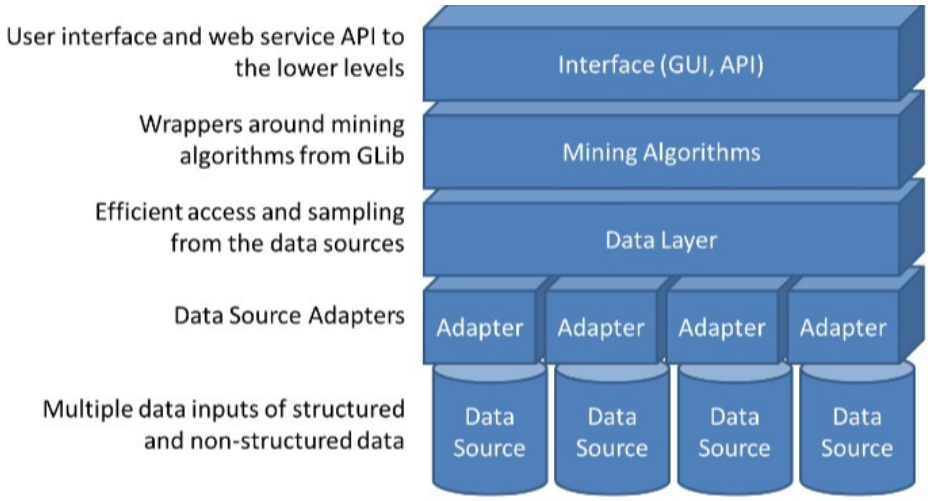
Figure 13 – Videk and its four main components.



Source: Kenda et al. (2013)

for interaction. As demonstrated in Figure 14, Videk is based on multi-layered and scalable architecture.

Figure 14 – Videk is based on multi-layered and scalable architecture.



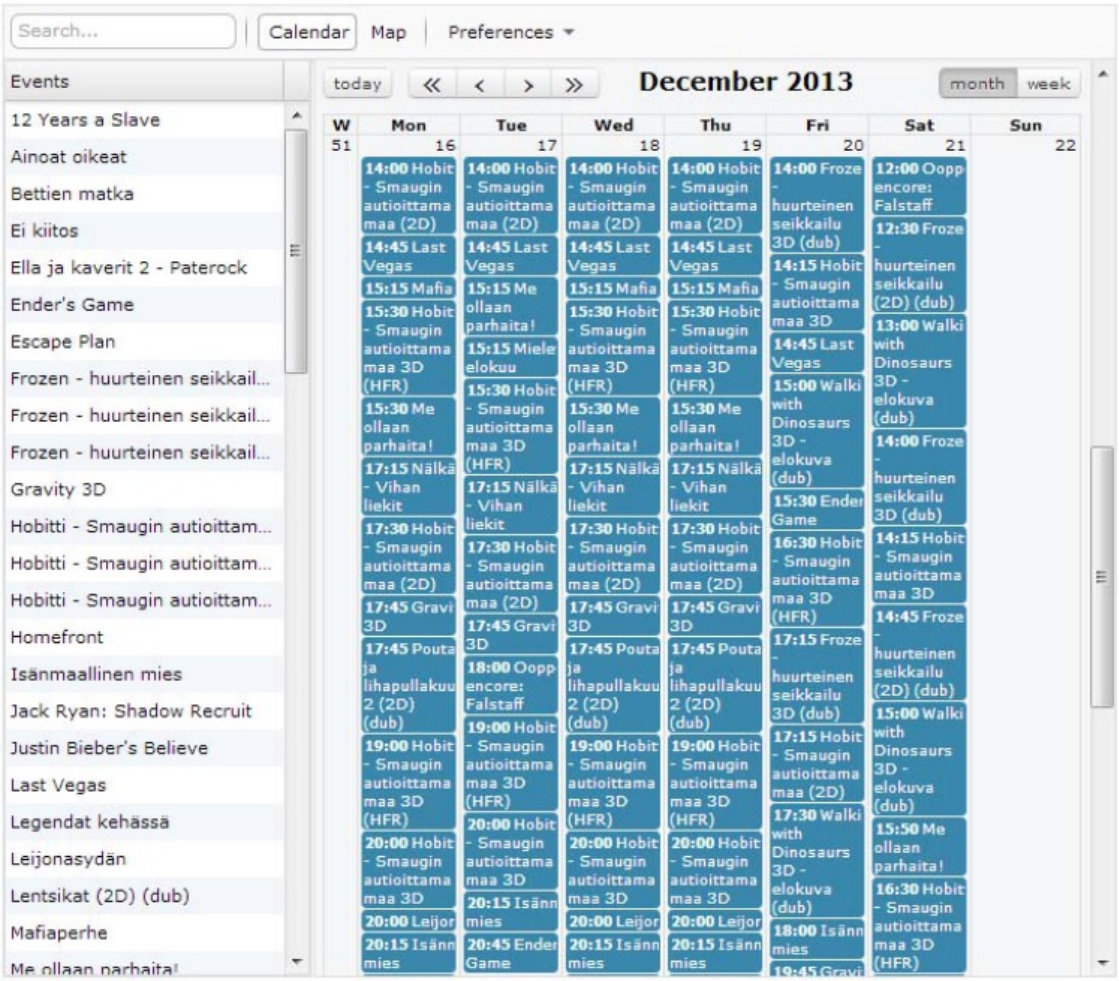
Source: Kenda et al. (2013)

As can be seen, this prototype provides information by aggregating sensor data and web data. However, Videk does not have automatic intelligence to extract information and deliver to the user. Moreover, it does not provide interaction, i.e. users cannot leave their opinions.

Rantanen e Sillberg (2014) developed an Event Calendar using open stream data

released by government bureaus and municipalities.

Figure 15 – Event calendar - system interface.



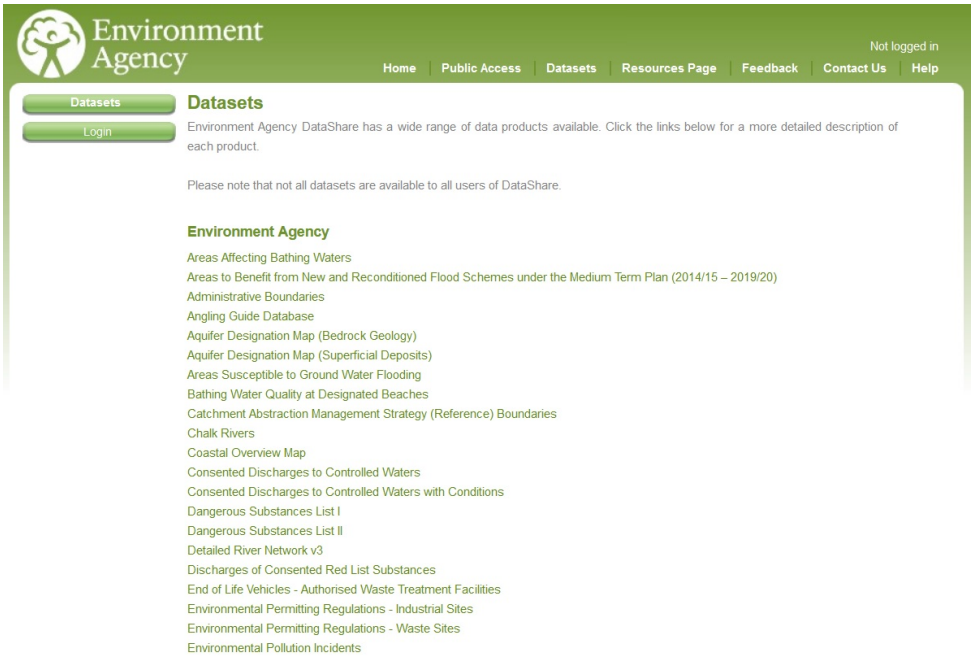
Source: Rantanen e Sillberg (2014)

These heterogenous data sources are used in order to increase the business opportunities associated with Open Data. They show how public collected data in the area of Satakunta, Finland, can be employed in an event calendar application. The system is composed of three different elements: client, front-end service and data storage. As presented in Figure 15, users can choose the options such as time, date, keywords and unrestricted text searches.

The application collects available dynamic data on the web and organizes it in a calendar structure. Users can interact with the application choosing their favourite events, however they cannot express opinions about it. Although the application uses dynamic stream data from web pages, it does not use any data direct from sensors. In this case, the application could use data from sensors and citizen in order to enrich the information.

Boyle, Yates e Yeatman (2013) investigated available data sources in the city of London. According to the authors, integrating data from diverse sources in the urban environment can provide an innovative vision and improve services in the cities. They state that recognizing and exploring the logical correlation among heterogeneous data are big challenges, but data aggregation across silos can be helpful. Moreover, new sensing systems will continue to present critical heterogeneity. Finally, the authors emphasize that multi-sector sensor data (coming from several Open Data initiatives led by government policy in the United Kingdom, as shown in Figure 16) has the power to revolutionise services. In summary, to deliver significant information of the virtual and physical world it is necessary to collect and fuse data from the main sources such as the Web, Sensors and Citizens.

Figure 16 – Open Data led by the government in the UK.



Source: Boyle, Yates e Yeatman (2013)

2 Characteristics and Issues of Heterogeneous Data Sources

2.1 Introduction

Recently, data is being produced and warehoused at unprecedented rates. In this chapter, individual data, sensor data stream and web data are investigated in order to identify their main characteristics and issues. In most cases, by knowing how to deal with such issues, it is possible to track problems and reduce errors.

2.2 Characteristics of data created by Individuals

Nowadays, sentiment analysis are becoming a very popular task. Due to the popularity of Web 2.0, a vast amount of data is being generated by individuals around the world. According to Devlin, Rogers e Myers (2012), Facebook users generate 2,7 billion likes per day, equally Twitters are creating 400 million new tweets by active users each day and Youtube is uploading 72 hours of video every minute. Under these circumstances, Twitter are generating more than 7 Terabytes (TB) of data every day and Facebook around 10 Terabytes (TB). There are several other web sites that offer to users the opportunity to leave comments and express opinions/sentiments about a topic. Thus, the scenario described above creates an ideal environment to investigate what people think and feel.

Sentiment analysis is also seen as opinion mining, sentiment mining, review mining, subjectivity analysis, appraisal extraction, sentiment classification or polarity classification. An opinion presents four main characteristics: topic, holder, claim and sentiment. Routray, Swain e Mishra (2013) state that: "The *Holder* relies on a *Claim* relative to a *Topic*, and usually correlates with the belief, a *Sentiment* such as "good" or "bad". For this reason, a number of researchers have emphasized that sentiments can be positive, negative or neutral (POIRIER et al., 2011), (BOLLEGALA; WEIR; CARROLL, 2013), (PANG; LEE, 2008), (RASTOGI; SINGHAL; KUMAR, 2014). To correlate the polarity to a sentence (positive, negative or neutral), it is necessary to count the number of words with positive semantic assimilation and negative semantic assimilation (POIRIER et al., 2011).

Sentiment Polarity Classification or *Polarity Classification* are binary classifications that put labels in opinionated documents (PANG; LEE, 2008). The principle of sentiment classification depends on the sentimental polarity of the opinion manifested in the text e.g. *"favourable"* or *"unfavourable"*, *"positive"* or *"negative"*, *"thumbs up"* or *"thumbs down"*.

Colbaugh e Glass (2013) state that it is important to evaluate opinion and sentiment toward a particular topic according to the following four points:

1. Sentiment (positive, negative or neutral);
2. Affective power / emotional force (pleasant, unpleasant);
3. Enthusiasm / arousal (excited, calm);
4. Rule / Dominance (dominating, dominated).

Sentiment analysis has been studied on four different levels:

1. As document level classification by (PANG; LEE, 2008);
2. As sentence level classification by (LIU; HU; CHENG, 2005);
3. As phrase level classification by (WILSON; WIEBE; HOFFMANN, 2005);
4. As word level classification by (KHAN; KHAN; KHAN, 2013).

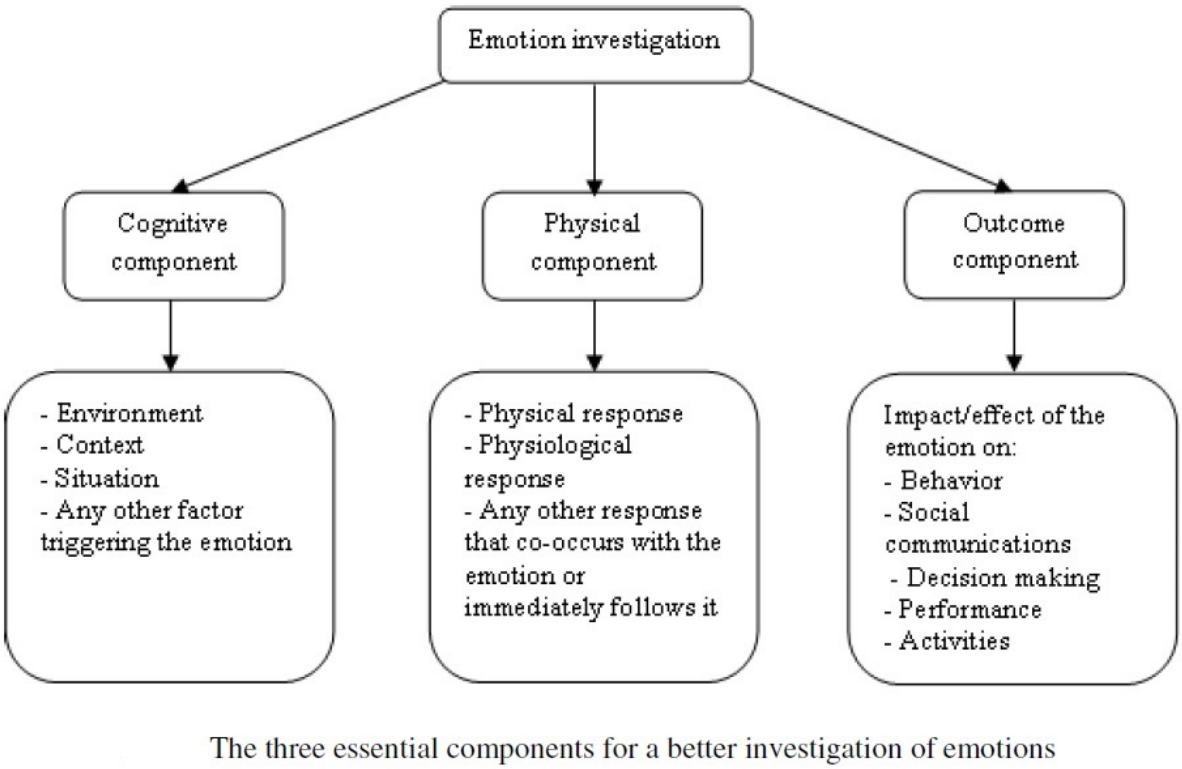
Similarly, Rastogi, Singhal e Kumar (2014) and Routray, Swain e Mishra (2013) highlighted four more characteristics:

1. Feature level (RASTOGI; SINGHAL; KUMAR, 2014);
2. Entity-level (RASTOGI; SINGHAL; KUMAR, 2014);
3. Opinion extraction (ROUTRAY; SWAIN; MISHRA, 2013);
4. Subjectivity classification (ROUTRAY; SWAIN; MISHRA, 2013).

Mousannif e Khalil (2014) indicate that sentiments (emotions) must be investigated in line with three important areas of investigation: cognitive, physical and the outcome. The cognitive considers the environment, the context, and situations that caused the emotion. The physical component first observes the physical reaction, e.g. facial expression, commentaries, postures, verbal or non verbal attitude. Secondly, it considers physiological reactions (pulse, blood pressure, respiration, etc), all these effects accompany the emotion or occur immediately after the event. Finally, the outcome component examines the impact of the emotion, for example, on attitude, social conversation, decision making, achievement, tasks and ventures. Figure 17 shows these branches in detail.

The following steps are highlighted by Rastogi, Singhal e Kumar (2014), aiming to obtain sentiments from given data. The first step is noise removal, which is necessary to avoid important data being processed together with irrelevant data like advertisements.

Figure 17 – Three areas that must be considered when investigating emotions.



Source: Mousannif e Khalil (2014)

Classification associates the data with the correct domain. Name Entity Recognition is necessary for recognizing the entity and the sentiment surrounding it. Subjective classification distinguishes between subjective or objective sentences. While a subjective sentence holds sentiments, an objective sentence presents facts and figures. Feature selection notes the features and the associated sentiment. Sentiment extraction can use several approaches to recognize sentiments like supervised or unsupervised learning and lexicon based method. Figure 18 shows a scheme of the Sentiments Analyses Steps.

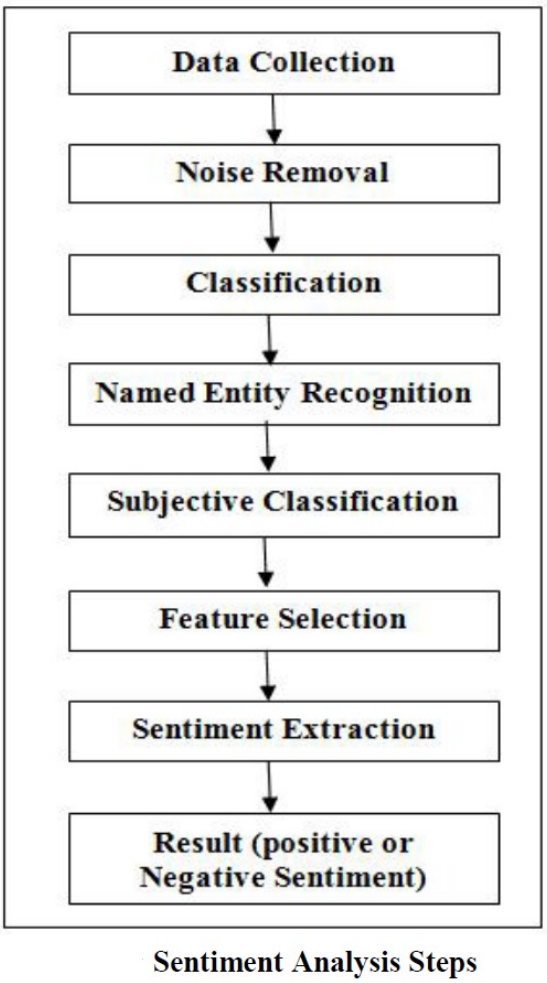
In addition, the authors have noted the following challenges:

Anaphora Resolution: is the use of a word (such as “it” or “do”) to avoid repetition of a preceding word or group of words. The issue here is to find what pronoun or a noun phrase it refers to.

Parsing: aims to describe a word from the point of view of classification, analysis of a sentence. The objective is to identify the subject and object of the sentence.

Sarcasm: The issue here is to distinguish when a word is expressing scorn or contempt and recognize ironical expressions.

Figure 18 – Steps to extract Sentiment Analyses.



Source: Rastogi, Singhal e Kumar (2014)

Poor grammar: Social networks present poor spelling and punctuation, lack of capitals, slang and several unrecognised abbreviations.

Ghag e Shah (2013) state that Sentiment Analysis is a text based interpretation issue, but some obstacles make it complicated when compared to conventional text based analysis. The authors listed the main problems for achieving the Sentiment Analysis: dealing with negations, words with numerous significance - polysemy, domain generalization, slang and many other challenges.

Dealing with negations: When a negation appears near an adjective, one can conclude that the polarity is the inverse of the polarity of that adjective. Simple language processing techniques and classic mathematical models on their own are unsuccessful when dealing with negation.

Words with numerous significance - Polysemy: It is a difficult task to

detect the correct significance of polysemy and its correct context.

Identifying Slangs: Casual dialect, jargon and slang are frequently used in online texts. Ex. n8 is a slang for night. These words are not listed in dictionaries but are very popular among young people. Consequently, the results of sentiment analysis can be improved if the slang is associated with the original word.

Domain Generalisation: In most cases, sentiment analysis is achieved focusing on a specific domain for better results. However, a generalised sentiment analyses still requires attention because a word or sentence with positive polarity in one domain may present negative polarity in a different domain.

Inclusion of different languages: Most sentiment analysers readily recognize English words, but only a few systems can identify Chinese and Vietnamese languages. In this case, the development of a Multilanguage opinion miner could provide a broad view of opinions and sentiments. This issue has been suggested by Routray, Swain e Mishra (2013)

Keeping Opinion for a period of time: Opinions can change over a period of time.

Hidden Sentiments Recognition: Recognising hidden sentiments like joy, disgust and anger is a difficult task.

Revision of Dictionaries: It is necessary to review words in the dictionary in order to include new words and remove words which are not being used. The accuracy of the dictionary hugely influences the efficiency of the sentiment analyser.

2.2.1 Denoising Individual data

Noise composed by individuals must be reduced. In this case, noise emerges from spam, bias and lies. In summary, it is important to identify what is noise and separate it from useful signal/information in order to provide accurate information.

2.3 Characteristics of Sensor data

IoT data are usually generated by a large volume of sensors. As a rule, sensors create data sequentially or when an event happens. Consequently, sensors produce data that need to be collected, analysed, aggregated or fused and interpreted. After all these steps, systems must be ready to provide representation, delivery and reaction (PERERA et al., 2014). In addition, not only sensors are producing stream data but also several other sources. There are several issues regarding IoT data:

1. Data streams can be defined as unlimited sequences of time-varying data elements on the assumption that recent information is more relevant as it describes the current state of a dynamic system (CORCHO, 2011);

2. Some versions of sensors do not have a sufficient quantity of energy and processing power to do overall data processing (SHO et al., 2008);

3. Sensors can face numerous technical obstacles and send data that is out of date, inaccurate and contradicting each other (YE; DOBSON; MCKEEVER, 2012).

4. To totally exploit sensor data it is essential to describe it as Linked Data (BIMSCHAS et al., 2011). However, as Yu e Liu (2011) state, most of the data linkages are done individually or are very disperse.

5. Barnaghi et al. (2012) state that sensory data represents environmental perception and need time, location and other explanatory attributes to make the data more significant. Moreover, Ye, Dobson e McKeever (2012) indicate that sensors deliver various types of data: binary, continuous numeric, and featured values.

6. Ye, Dobson e McKeever (2012) emphasize that each data type must have a specific technique to analyse the data.

7. According to Perera et al. (2014) the IoT group together various technologies to complete its purpose (e.g. sensor, hardware/firmware, semantic, cloud, data modelling, storing, reasoning, processing, distributing). As a result, the data has interoperability issues.

8. Boyle, Yates e Yeatman (2013) point out that data vary from rigid datasets to active data-generating network embedded sensing systems, social media streams, participatory sensing systems, and probably, include various modelling capabilities. Under these circumstances, the data have heterogeneity issues too.

9. "Heterogeneous applications can lead to a challenging Veracity problem (SHULL, 2013)" . Consequently, IoT data can vary in quality and/or can be incomplete.

10. It has been highlighted that data saved at specific sensor networks are isolated and cannot readily be included. Raw data may have to be aggregated or derived (ZHANG; TJHUNG, 1999). Similarly, Perera et al. (2014) emphasize that sensor data fusion is a technique of joining sensor data from various sensors to obtain more precise, more complete, and more reliable information that could not be produced through a single sensor. In brief, fusion and integration are mandatory and open questions.

11. Data can be structured, unstructured or semi-structured and have to be integrated or fused in order to glean a compact synopsis and semantically rich metadata (ZERFOS et al., 2013). Some examples of each data type are:

    a) Unstructured Data: audio and video streams, analog data, GPS, tracking information, text and media files;

    b) Semi-structured data: headers, web pages, XML, EDI, email.

    c) Structured Data: billing records, data warehouses, databases.

12. Zerfos et al. (2013) identify four stages to process stream data:

    a) Data collection stage;

    b) Data filtering stage;

    c) Data analyses stage;

    d) Data consumption stage.

13. Finally, as researchers have noted, security, confidentiality and privacy may have to be observed as well (RANTANEN; SILLBERG, 2014).

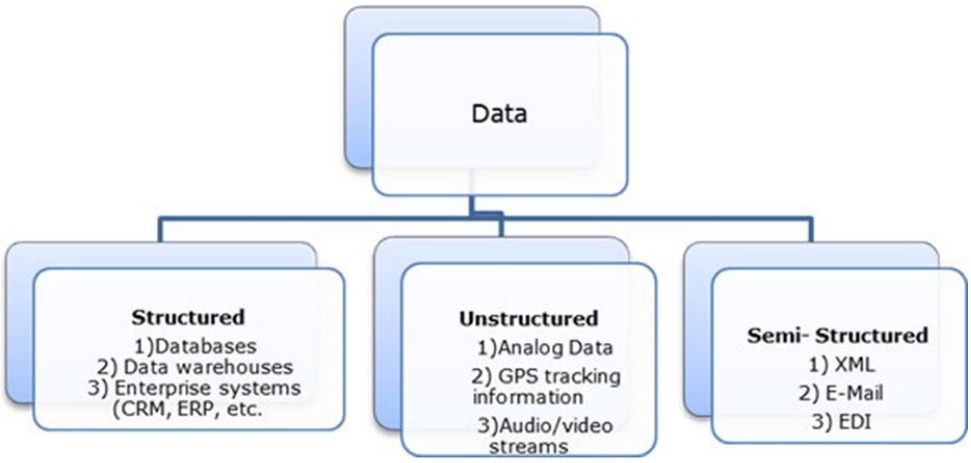
2.3.1 Denoising sensor data using Discrete Wavelets transform

In the case of sensor data, an important procedure is the removal of noise. The noise that come along with data, disrupt the process at all stages. Thus, it is necessary to remove the noise in order to optimize the accuracy of the information. According to Berkner e Wells (1998) wavelet coefficients of a function are commonly large in regions where a function is irregular and small in smooth regions. If a function is polluted by noise, this noise influences the wavelets at small scales. Under these circumstances, the biggest part of those coefficients contain the noisy part of the signal and only a minority part of coefficients is related to strong singularities in the basic function. The right threshold scheme eliminates most of the noise and retains the large coefficients. A smooth approximation of the function f is retrieved via the inverse wavelet transform.

2.4 Characteristics of Web data

Web data often comes from purchase transactions, web traffic, rewards programs, quarterly business reports, Twitter, Facebook, blog content, etc. These sources generate e-mails, documents, social media forums, web pages, web log files (including click-stream data), search indexes, etc. In the same way, the Internet of Things spreads lots of devices which are generating data derived from sensors attached to machines, and products / devices that collect all kinds of data, from GPS coordinates, temperature data and usage data.

Figure 19 – Types of Data.



Source: Sheth e Thirunarayan (2012)

Figure 19 shows that data can be classified as structured, semi-structured, unstructured. Sheth e Thirunarayan (2012) provided definitions for the different types of data:

Unstructured Data: In general, this type of data does not have any predefined structure established by their topic. It is usually classified as Grammatical Text or User-Generated Content.

Semi-structured Data: In most cases, this includes tags or other markers to restrain and catch hierarchical and fused structures associated with semantic labels.

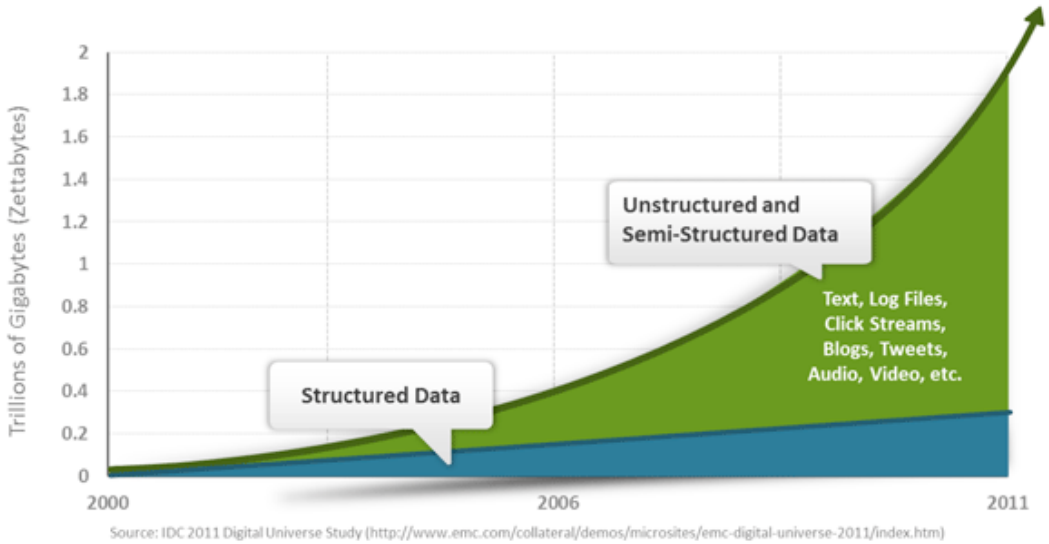
Structured Data: As can be expected, this type of data has well-defined formal syntax and is connected with a data model.

Figure 20 shows that the quantity of unstructured data is growing exponentially. However, most part of the systems are able to manage only structured data. Therefore, the challenge now is to provide enough intelligence to the systems, so that they work with all types of data.

The main challenge is to find the path from data to foresight, in other words how to classify data and learn how to mine simultaneously among different types and sources of data. From the above it is possible to conclude that systems must be ready to collect, analyse, mine and extract insight among all different types of data.

The scenario described in this section suggests that future research should focus on developing combined techniques to analyse raw data in several types: structured,

Figure 20 – The quantity of unstructured data is growing rapidly.



Source: IDC - Digital Universe Study (2011)

semi-structured and unstructured data. These models can be designed to extract various important information about the virtual and physical world for an interconnected society.

3 Software Development Lifecycle

The Software Engineering area has defined the software development process very well. If all recommendations are followed strictly, a software development process has a good probability of success. Therefore, the platform was designed using established Software Engineering methodologies.

3.1 The overall description

Nowadays, people are surrounded by data and are constantly producing new data. Never before have they had so many sharers/devices working together and producing such massive amounts of data. However, few applications are able to provide the basis for individually and socially aware applications and services. So, the vision that drives this research is: 1) Collect correlated data from various sources (data sources are described in the Figure 21); 2) Aggregate the data; 3) Apply different techniques, that can provide the platform with an understanding about the individual, his environment (society) and health, his sentiments and topics (or events) that matter to him.

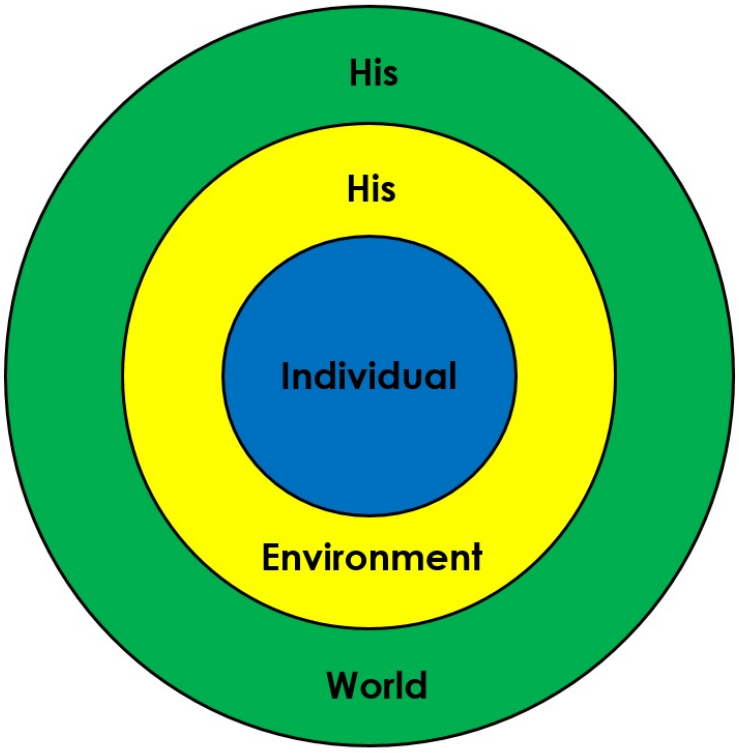
The platform will be divided into three modules. Each module will collect data from a bunch of different sources. This data can be structured, semi-structured or unstructured. Thus, each module must be analysed separately and designed in order to extract information correctly. The output data of the platform will be the entry data for individual and social aware applications and services.

3.2 Individual data at the heart of the Design

According to Fauquex et al. (2015), it is difficult to find techniques that properly insert the end-user needs into the IoT development process. On the other hand, Knoeri, Steinberger e Roelich (2016) state that is particularly important to consider the end-user, and their demand, as the fundamental element of the project. In addition, the authors say that end-users request infrastructure and services that are ready to satisfy their particular needs and wants. The picture presented by Figure 21 has the end-user (or the previously mentioned Individual) at the heart of the role platform. Individual data is the most important data, and correlated data (data from sensors and the web) must be aggregated to Individual data.

For these reasons, placed at the center of the Figure 21 is Individual data, meaning that the information delivered by the platform will be produced in accordance with the

Figure 21 – Individual data is the main driver. Correlated data must be aggregated taking into consideration individual data first.



Source: The author.

Individual's perspective. In the second layer is placed sensor data. This data is generated by sensor devices located around the individual (these sensors could be wearable devices). Sensor data must be correctly correlated with Individual data, delivering information about the environment in which the Individual is situated (or even related to Individual's health). The last layer is web data. It must be able to bring additional information that is interesting to the Individual or is correlated with the Individual's data.

This design has the objective of answering the first question of this research at the section 0.4.2.

3.2.1 Module 1 - Individual Data - General Requirements Specification

1. **Overall description:** - This module is responsible for collecting, transforming, storing and analysing personal data that is available in social media, such as Facebook, Twitter, LinkedIn, and others. It has to combine data from several sources, detect individual mood, learn behaviour and preferences, constantly update values, store, and send this information to the next stage. Semantics can also play a role here. According to Mika, Elfring e Groenewegen (2006) implementing the semantics

of social network data is essential for the fusion of social network information, particularly in a heterogeneous environment where the specific sources of data are under diverse control. (The first prototype will use only Twitter, because Twitter uses a small set of words that can easily be analysed).

2. **Specific Requirements:** - Each user must be uniquely identified by the software. Thus, the software will identify the user asking him to connect the application with his social networks; then the software will track his posts, registering day, hour and location (if is available), recognize the topic, identify the mood. The software will follow the traditional process: collect, transform, store, analyse. **Users and Characteristics:** - Data will differ according to the user. Each user must be analysed separately and their personal data collection properly stored.

3. **Functional requirements:** - Fundamental actions that must take place in the software:

   Collect: there are several techniques that can be used for data collection, for example, website crawling, also known as site scraping or web harvesting. Moreover, each social media website has made available its own API, which makes the data collection process easier. The first prototype will use only Twitter and an API, called Twitter4J, from a third party service. This micro-blogging service has a very significant number of users who can write various short text messages, known as tweets, about any topic and event (tweets must be up to 140 character long) (PISKORSKI; TANEV; BALAHUR, 2013). Other social network services, like Facebook, accept user entries that can be enormous. Moreover, according to Tan et al. (2014) Twitter has found an effective way to expose public opinion in a timely manner, which is useful for this platform. Therefore, Twitter profiles of several users were collected at regular time intervals. An investigation using geo-tagged tweets (MITCHELL et al., 2013) and the classification scheme using hashtags will be investigated in future work.

   Transform: Each user account collected by the Twitter4J API is transformed into a .txt file. After that, a technique called Term Frequency – Inverse Document Frequency (TF-IDF) is applied. TF-IDF is a technique used in the field of natural language processing (NLP) for identifying those words with importance and those words not relevant (ZHANG; YOSHIDA; TANG, 2011) (TRSTENJAK; MIKAC; DONKO, 2014). TF counts the number of repetitions of each word in the document, IDF sets to zero (or near zero) the count of those words which occur in many documents. Thus, using TF-IDF method, each word in the .txt file is assigned a value (or weight) that is its TF-IDF score. Each Twitter user's profile is transformed in a vector of numbers that is stored in the Excel file. Finally, this step can also include removal of noise (smoothing), aggregation, generalization and normalization.

Store: MySQL or NoSQL will be the database used to store and retrieve information. MS Excel is used during the prototype development.

Analyse: An exploratory analysis using Self-Organizing Maps (SOMs) will be applied to the sets of vectors containing the TF-IDF scores of the words. Basically, SOMs will display each vector of scores to the network, bringing together similar data weights with similar neurons.

Visualise: Visualisation is very important when processing investigative data analysis (MURTAGH; HERNANDEZ-PAJARES, 1995). Interfaces must be designed to adjust themselves for different devices. These interfaces will be designed according to the rules published by this author during her Masters Degree (FARIA, 2007).

4. **Implementation Constraints and Design:** - Interfaces to be designed: user interfaces; hardware interfaces; software interfaces and communication interfaces.

5. **Unit tests:** Unit test: refine, review and update.

6. **User Documentation:** - user documentation will consist of simple explanations and instructions for operating the system.

3.2.2 Module 2 - Sensor Data - General Requirements Specification

1. **Overall description** To better understand the context of this subsection it is important to highlight five characteristics of sensor data:

   a) Data streams can be defined as unlimited sequences of time-varying data, in which recent information may be more relevant, as it describes the current state of a dynamic system (CORCHO, 2011).

   b) Some versions of sensors do not have a plentiful supply of energy and processing power to perform comprehensive data processing (SHO et al., 2008).

   c) Sensors can face numerous technical obstacles and send out-of-date, inaccurate and contradictory data (COSTA; CUGNASCA, 2010).

   d) IBM has noticed that "heterogeneous applications can lead to a Veracity problem" (DEVLIN; ROGERS; MYERS, 2012). Consequently, data can vary in quality and/or can be incomplete.

   e) Determining the quality, validity, and trustworthiness of data are priorities in the IoT research area (LI et al., 2002).

   The first characteristic implies that sensors sense the physical world and send the data in the form of a signal. Signal is a function delineated according to an independent variable. This variable is usually time, but could express any number of features.

Nowadays, sensors are being placed in many different environments. So the sensor signal can suffer significant interference when modifying that variable and generating a signal with noise.

The second and third characteristics consider the fragility of sensors in terms of hardware. In general, sensors are small devices with several limitations. For example, they may have a limited amount of energy, low processing power, face technical obstacles, etc. These characteristics can result in generation of a signal with noise. Berkner e Wells (1998) highlight that in many applications in signal and image processing the observed data is influenced by noise. According to them, an important question here is: ''How can the underlying clear signal be separated from the noise? The appearance of data can be changed if the sample has a significant amount of noise.

Regarding the fourth characteristic, where heterogeneous applications can lead to a Veracity problem: if a system analyses (or fuses) uncorrelated data (noisy data), this can lead to incorrect information or bad predictions.

Finally, the fifth characteristic is essential for better understanding the phenomena captured by sensors and transmitted as a signal. A good quality signal (without noise) leads to trustworthy data, and consequently, the next step can produce better results or predictions.

In this research the main contributions made in the field of sensor data are summarised. Thus, we propose an innovative technique to improve the signal of a sensor, enabling a more trustworthy extraction of data. The proposed technique is based on the Discrete Wavelet Transform, which aims to clear the wavelet coefficients that contribute most to the noise than for the signal. It is not effective or possible for all signals because the signal must have favourable spectral characteristics, i.e. the information regarding the signal must be concentrated in a small number of wavelet coefficients.

2. **Specific Requirements:** First: To perform an exploratory data analysis of the signals to verify if they have the adequate characteristics suitable for applying a wavelet based denoising technique.

   Second: if the signal has suitable spectral characteristics, an investigation into the sensor signal using Discrete Wavelet Transform (DWT) must be performed. This technique may drastically reduce the volume of the data, thus enabling construction of a new compact representation of the signal. This technique is also useful for removing the noise.

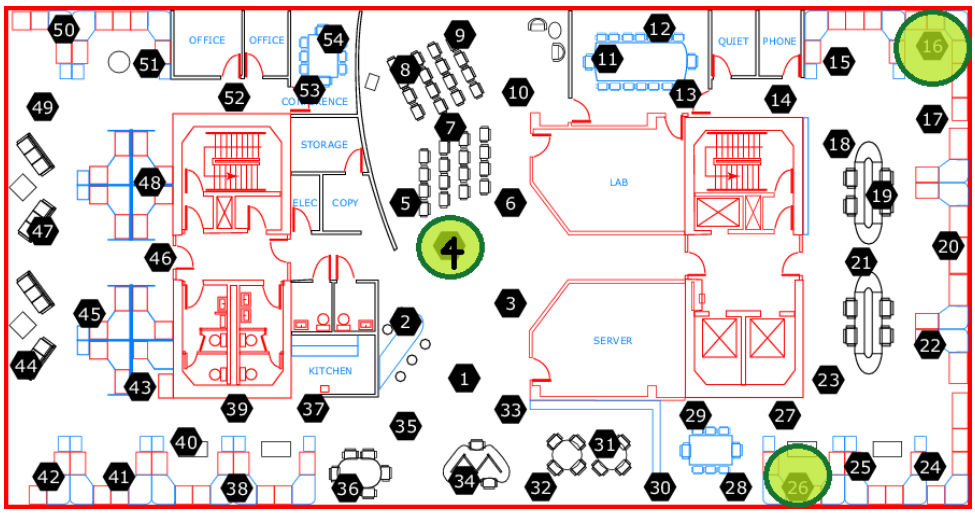
   Third: An application must be developed to show that the new reconstructed and denoised signal not only is useful, but can also improve the performance of applications.

Fourth: When a sensor signal is cleared of noise, it is easier to predict because all coefficients that do not contribute have been eliminated. Therefore, it is possible to know the specific signal of each particular measurement. Thus, each measurement performed by the sensor must be analysed and identified as a particular type of signal and stored with its associated semantic value.

3. **Functional Requirements:**

Collect: For the purpose of our experiment, a collection of sensor data from the testbed in the Centre of Communication System Research at the University of Surrey was investigated (NATI et al., 2013). The sample collected contains 16,384 data items. The data was obtained over two days from sensor nodes deployed in the offices collecting information about light levels. For the purpose of a second experiment, a collection of sensor data from the testbed in the Intel Berkeley Research Lab was investigated. This sample contains data generated by three sensors: (i) Sensor 4 - positioned in the center of the room, (ii) Sensor 16 - positioned in the upper corner, and (iii) Sensor 26 - positioned on the opposite corner to the sensor 16. Figure 22 shows the position of each sensor and highlighted in green are sensors 4, 16 and 26. These three sensor nodes collected measurements of temperature, humidity and light intensity.

Figure 22 – The position of the sensors 4, 16 and 26 are highlighted in green.



Source: Intel (2016)

Transform: According to Pietropaoli, Dominici e Weis (2013), solutions learning directly from raw sensor data are very susceptible to the quality of the training set. For this reason, a technique has been employed in order to transform raw sensor measurements (or raw sensor data). If the sensor signal has favourable spectral characteristics (i.e. most information conducted by the sensor signal is concentrated

in small number of coefficients) DWT is the best technique to denoise the sensor signal. In this circumstance, it is possible to set as equal to zero those wavelet coefficients that contribute most to the noise and least to the signal. Therefore, it is possible to reduce signal noise using discrete wavelet transform, first decomposing the signal in levels; second, generating vectors of wavelet coefficients; third, choosing the most energetic coefficients to reconstruct a new signal using Haar as the mother wavelet.

Store: After transforming each measurement as vectors, i.e. vectors that contain the most energetic wavelet coefficients (that are also denoised) each vector must be stored and that particular measurement must be semantically identified.

Analyse - exploratory data analysis and predictions: It is necessary to perform an exploratory data analysis of the signals to verify if they have suitable characteristics for undergoing a wavelet based denoising technique. To make predictions using a sensor signal it is necessary to do some basic analysis utilising Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF), Variance and Periodogram. Although these techniques are very basic, they are essential for extracting several characteristics of the signal in order to get a better insight into its structure. Careful analysis of these graphs makes it possible to determine the mathematical structure behind the time series, thus indicating the best way to model the signal and consequently to develop forecasting tools.

Visualise: Using virtual sensors is possible to visualise combined sensor data, i.e. data coming from different sensors. Moreover, by doing simple calculations, like averaging or finding the maximum/minimum values, it is possible to merge sensor data coming from multiple sensors that observe the same physical phenomenon (PIETROPAOLI; DOMINICI; WEIS, 2013).

4. **Implementation Constraints and Design:** DWT may drastically reduce the volume of the data, making possible the reconstruction of a new compact representation of the signal. This technique is also useful for removing the noise. After applying the DWT technique and removing the noise, each particular signal (that represents a specific measurement) is transformed into a vector of wavelet coefficients and stored with its proper semantic label. The stored vectors can be used for forecasting purposes.

5. **Unit tests:** Tests aim to show that the filtered signal is helpful for developing applications. In other words, it is useful not only for reduction in the numbers of coefficient wavelets, but also to reduce the number of errors that occur in applications.

6. **User Documentation:** Again, user documentation will consist of simple explanations and instructions for operating the system.

3.2.3 Module 3 - Web Data - General Requirements Specification

1. **Overall description** Answering questions posed by individuals in their natural language is a very challenging task for systems to undertake. Nowadays, when people are looking for answers on the web, using popular search engines, they receive a list of documents deemed relevant to the question. It would be better if they received a list of short and precise answers, thus saving time exploring several sources of information. Many researchers (KUZNETSOV; MOCHALOV; MOCHALOVA, 2016) (BHOIR; POTEY, 2014) (COMAS; TURMO; MàRQUEZ, 2012) are working hard to improve systems, making them capable of answering questions in order to assist individuals in their tasks. Some systems such as Apple Siri, Microsoft Cortana, Google Now, Amazon Echo, S voice, Blackberry Assistant, Wolfram Alpha and others, can make conversations, create reminders, make calls to contacts, search the web, etc. It is likely that a Question Answering System (QAS) would improve the experience of individuals and attract their attention to applications. For this reason, a Question Answering is the system chosen for the platform.

2. **Specific Requirements:** First, the system must be able to understand what is expressed in a question and how to harvest this information. Second, the system has to identify linguistic regularities and features in the question. Third, the system will use basic Natural Language Processing (NLP) tools (e.g. syntactic parsing, Named Entity Recognition (NER), etc).

3. **Functional Requirements:** The QA system will be divided into seven components or modules. These modules are responsible for:

    a) Entity Search: Matches query terms to dataset entities. Index/search temporal performance. Needs to support semantic approximations e.g. coping with different lexical expressions, abstraction levels. Will use thesauri and distributional semantics based approaches for semantic matching.

    b) User Interface (UI) and QA Pipeline: Integration of the QA components. Development of the Web interface for the QA system. Exploration of simple user feedback mechanisms (e.g. entity disambiguation).

    c) Question Analysis: Input: Natural language question. Output: Parsed question. Candidate entities and associated types. Candidate relations between entities. Lexical answer type. Candidate database operations

    d) Graph Extraction: Extract entities and relations from Wikipedia text. Preserving contextual information. Persist them as RDF graphs. Focus on fact extraction.

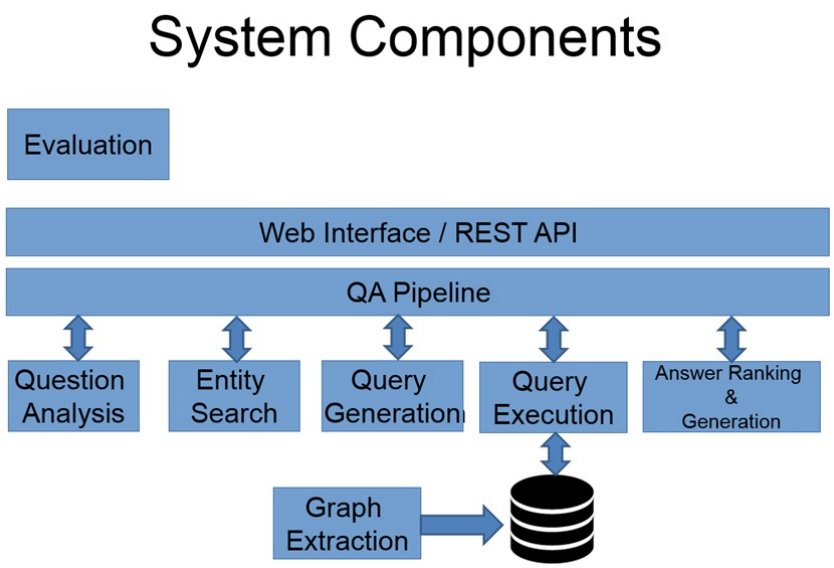
e) Query Execution / Answer Ranking and Generation: Input: Possible SPARQL queries. Output: Result sets. Ranking models and heuristic models for classifying the answers in relation to a question. Transform results in triple format to a natural language form.

f) Query Generation: Transforms the natural language query into a query in a logical form. Involves the interface between natural language and knowledge representation / logical models. Relation identification / extraction.

g) Evaluation: Automatic evaluation for the QA system using the Question Answering over Linked Data Test Collection (QALD-4).

4. **Implementation Constraints and Design:** The design of the system follows the scheme presented in Figure 23.

Figure 23 – Question answering system design.



Source: The author.

5. **Unit tests:** Using the latest QALD version, build a tool to calculate precision, recall and f1-measure for the example queries.

6. **User Documentation:** has the purpose of helping users to pose questions correctly and informing them how to proceed in case of bugs.

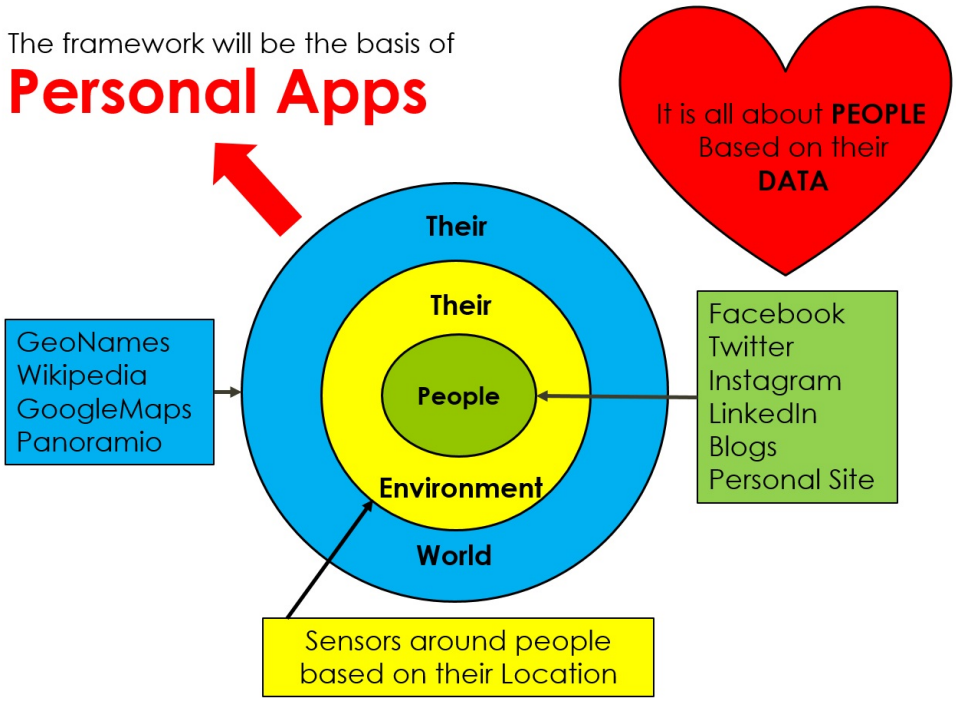
3.2.4 Module 4 - Process of Integration

The platform proposed in this research has a particular approach to a specific objective that is focused on the user, his activities and interests. This vision is described in Figure 24. To achieve the intended purpose, the platform will provide the overall structure

to interrelate all components described in this chapter in a way that allows them to work together. It will be a web-based layered structure capable of combining tested softwares into a integrated whole.

The process of integration has the challenge of assembling all elements, making each component of the system function properly as a whole, reach the design properties, specify interfaces, while performing the verification and validation of each part. As can be expected, these tasks have a big demand for different expertises, robust infrastructure of hardware and modern platforms for software development. For this reason, this part of the project was submitted to Fapesp Pipe[1], in order to obtain the necessary resources to accomplish these tasks.

Figure 24 – A vision that drives this research.



Source: The author.

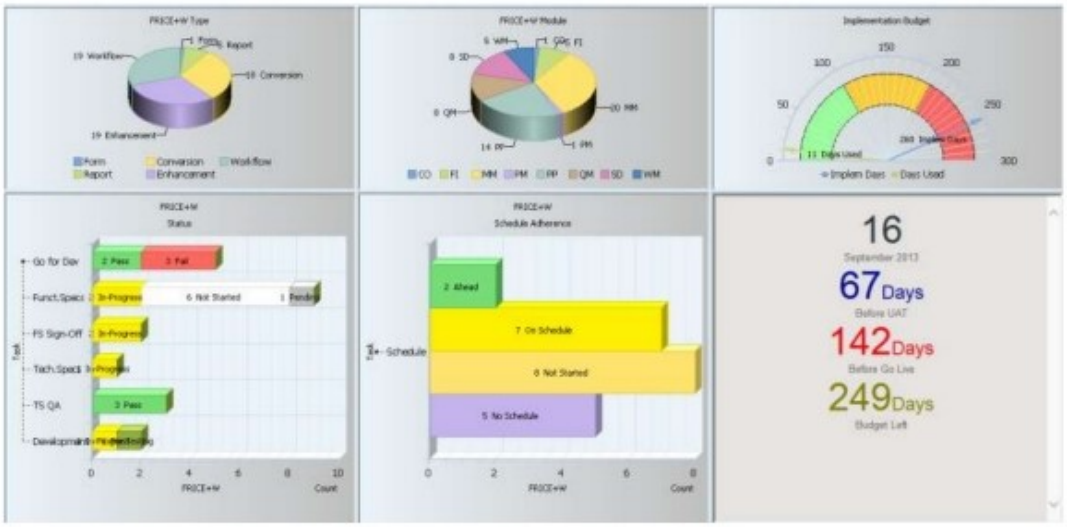
3.2.5 Module 5 - Applications: Mobile Applications and Dashboards

Mobile application, or App, is a technology relatively mature. The best way to reach the final user (or Individual) is developing a mobile App for smartphones. For this reason, the output data from the platform will be the basis of various mobile App (for each scenario will be developed an App) . These Apps will be released as free downloadable technology.

[1] Sponsored by PIPE Fapesp Fase I <http://www.fapesp.br/pipe>

On the other hand, the manipulation and analysis of large datasets for authorities and managers can be done through visual analytics tool, or dashboards. These tools are useful to help users interpret the message behind large scale datasets. Dashboards can provide data understandability as shown in Figure 25.

Figure 25 – An example of a web dashboard report.

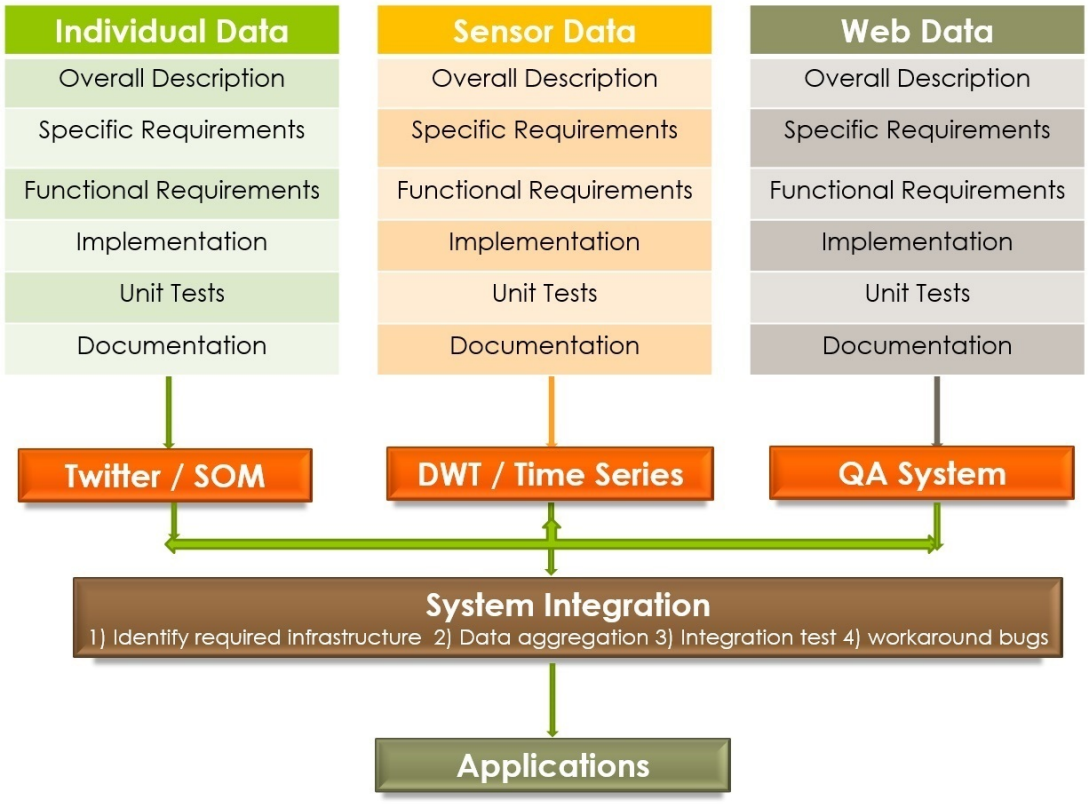


Source: Google Images

4 Development & Verification

Figure 26 is a summary of Chapter 3, where each source of data was separately analysed and presents its respective output. This output is the input of the platform. The platform requires a system integration and an infrastructure capable of dealing with massive amounts of data. Data aggregation involves fusion of correlated data. Tests of accuracy, interfaces, security and privacy must be performed and bugs removed. The output of the platform is the input of applications, i.e. mobile applications (for individuals) and dashboards of data (for authorities and managers).

Figure 26 – This figure represents a summary of Chapter 3.



Source: The author.

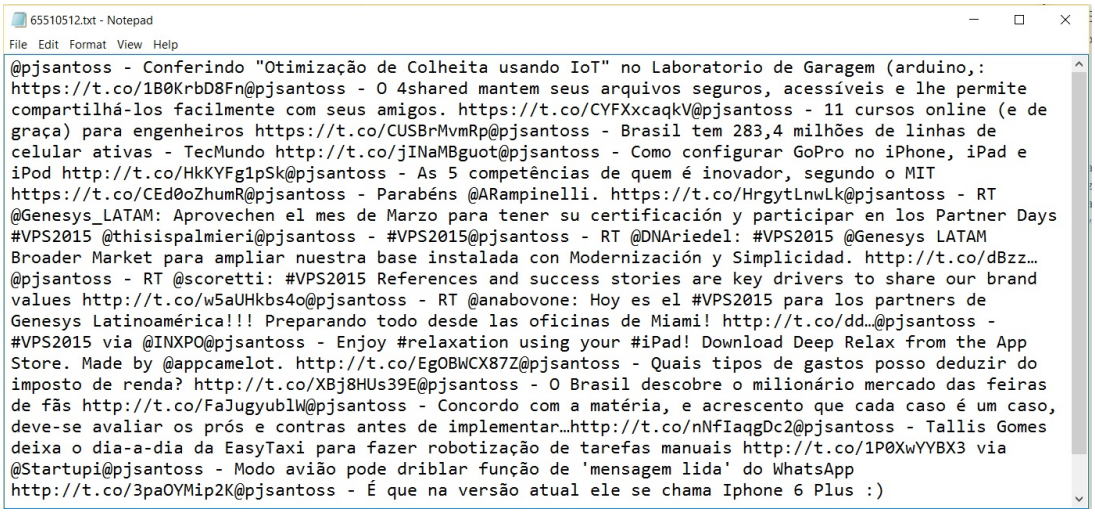
The following sections describe the results obtained from each experiment. There are many techniques capable of dealing with these data sources. Careful research presented in Chapter one and two led to the techniques described here. Each technique produced the desired result, indicating that the outlook for the future is good.

4.1 Individual Data - Twitter and Self-Organizing Maps

Data acquisition from the Twitter social network is made via Twitter4J API, which is a library in Java format that can be freely downloaded from the internet. This library is not considered official, but it successfully obtained the required data when collecting users' profiles and all words posted on their timelines.

The Twitter4J API needs a user account ID to be provided with the request. All tweets posted to this user's account are extracted and transposed to a .txt file. The API is also able to read the names of the followers and get the user's entire network of contacts. Subsequently, the profiles of the followers also will be analysed and have their tweets collected in .txt files. Finally, a scan is performed to remove unnecessary characters such as hyphens and links. The tags of the html language are removed by Twitter4j. Figure 27 shows the result obtained after extracting the profile. The name of the file is the user account ID. More than 700 profiles from Twitter were collected using this methodology.

Figure 27 – The created .txt files were standardized as: @usuario: - tweets.



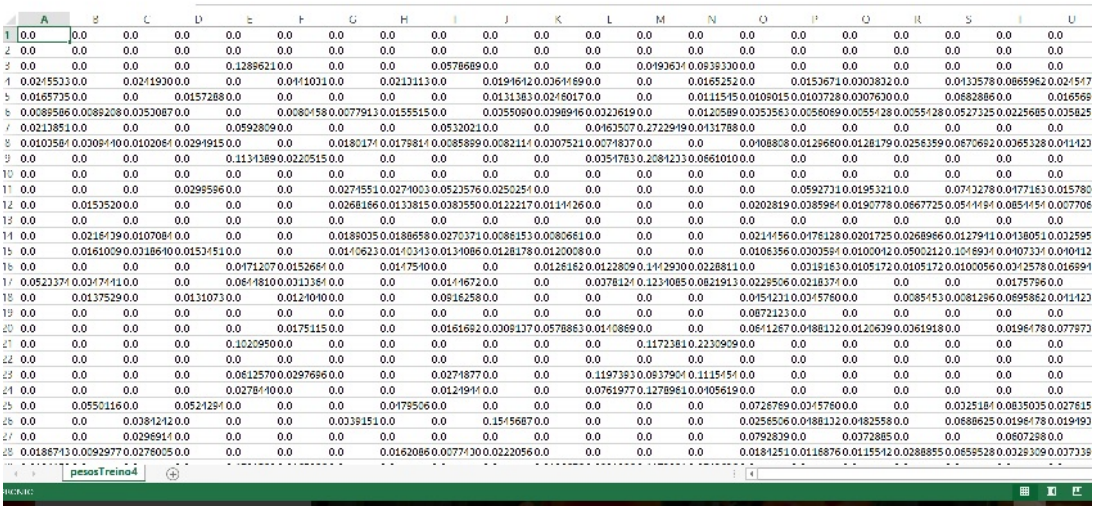
Source: The author.

Next, it is necessary to convert the .txt files into numerical data, since the Self-Organising MAP (SOM) is able to read only numerical values. A leading method for doing this conversion is called Term Frequency (TF) Inverse Document Frequency (IDF). This technique is also known as TF-IDF.

Before executing TF-IDF, it is necessary to specify certain rules about unnecessary and high-frequency words, as conjugations of verbs and trivial articles. The least significant words have been smoothed out or set to zero, leaving the most expressive words with high values. By doing this, the least significant words are removed first.

Each .txt file is transformed into a vector of values, which can be converted to

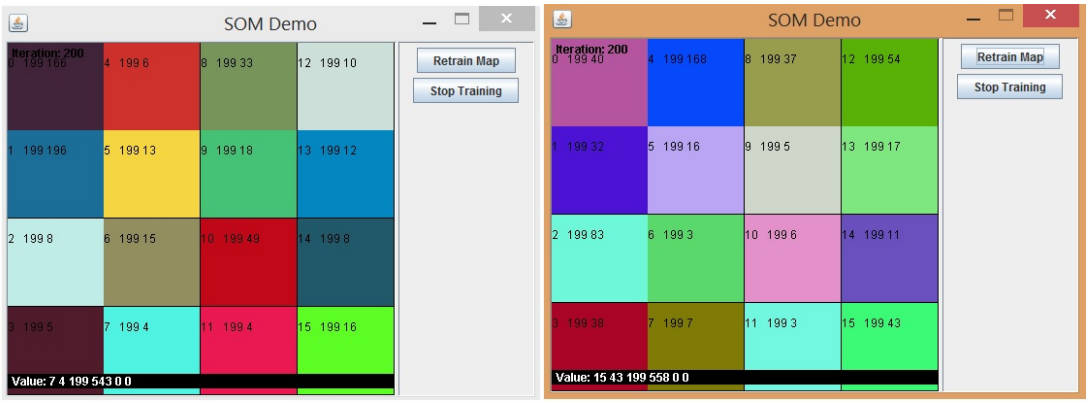
Figure 28 – Each word has a numerical value (or weight) according to its frequency.



Source: The author.

become a line of a spreadsheet. In this case, the data is transferred to Microsoft Excel software and the final file is saved in .csv format that is compatible with the request of the SOM emulator. Figure 28 shows some values that are 0. These are for words like no, na, em, para, de, etc. There are also values greater than 0. These words have a numerical value (or weight) that corresponds to their frequency.

Figure 29 – Results presented by SOM algorithm. Each node has three numbers: the ID of the node, the number of iterations (or training) and the number of profiles in each module



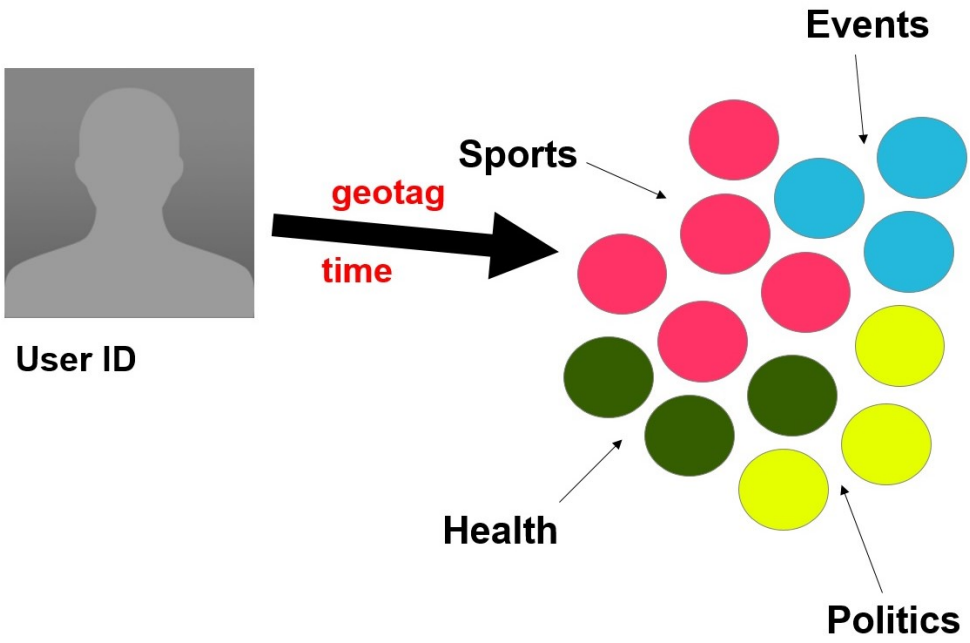
Source: The author.

SOM is capable of providing meaningful maps that help to explain complicated datasets using a visual approach. Consequently, SOM was applied to the data set so that similar profiles are clustered. For the simulation of the algorithm, some code in C# named

SOM, was downloaded, free of charge, from the website www.mql5.com/en/articles/283 in October 2015. After training, the results displayed in Figure 29 were obtained.

The purpose of this methodology is to group similar profiles of individuals, i.e. each individual can be grouped according to the words found in his or her profile. Moreover, each individual can be identified by his or her user account ID. When grouped, these individuals can receive related information that matters to them. As shown in Figure 30 one specific individual will be directed to one specific group where are several other individuals with similar profiles. Thus, it is possible take actions that are useful for that group.

Figure 30 – Each word has a weight according to its frequency.



Source: The author.

As future work, we intend extend the methodology to include data such as geotag and time.

4.2 Sensor Data: Denoising sensor signal using Discrete Wavelet Transform

One of the possible techniques for reducing the noise at the receiving side of a communications system employs the wavelet transform (MOMEN; AHMADI-NOUBARI; MIRZAEE, 2006). The typical noise reduction method based on wavelet transform is called denoising is used, for example, to reduce noise in images.

The wavelet transform is a representation in the plane time-frequency. Any signal can be completely recovered by means of the wavelet coefficients and the inverse wavelet

transform. An approximation of the signal can be obtained using a subset of the wavelet coefficients.

If, for example, the received signal has a reasonable Signal-to-Noise Ratio (SNR) but is not good enough to achieve the desired Bit Error Rate (BER), then the elimination of the wavelet coefficients that are less than a given threshold may result in an improved SNR, because the small coefficients carry more information about the noise than the desired signal. On the other hand, for certain signals the energy is concentrated only in one part of the spectrum. Thus, if only the wavelet coefficients corresponding to this part of the spectrum are used, the SNR also improves (PERCIVAL; WALDEN, 2000).

According to Berkner e Wells (1998), wavelet coefficients of a function are commonly large in regions where a function is irregular and small in smooth regions. If a function is polluted by noise, this noise influences the wavelets at small scales. Under these circumstances, the greater part of those coefficients contain the noisy part of the signal and only a small part of the coefficients are related to strong singularities in the basic function. The right thresholding scheme eliminates most of the noise and retains the large coefficients. A smooth approximation of the function is retrieved via the inverse wavelet transform.

4.2.1 Methodology of the Experiment

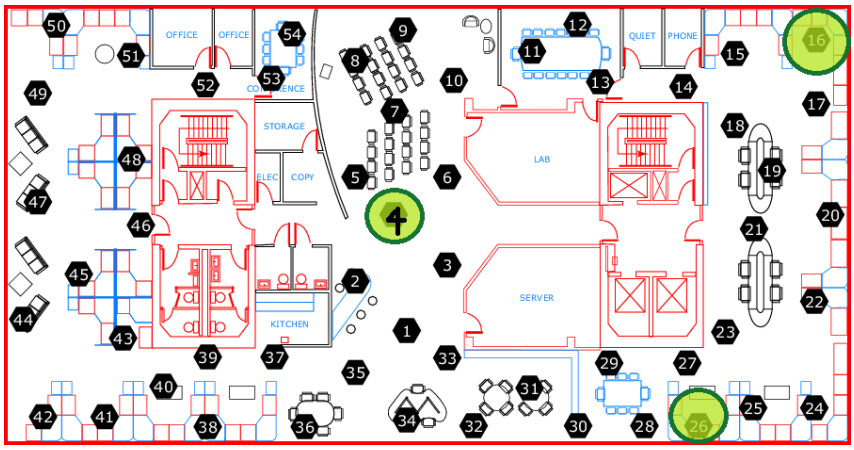
The prime objective of this section is to demonstrate that DWT is the ideal technique to denoise the sensor signal if its signal has favourable spectral characteristics. This means that, the main information conducted by the sensor signal is concentrated in small number of coefficients. Under this circumstance, it is possible to set equal to zero those wavelet coefficients that contribute most to the noise than for the signal.

For the purpose of our experiment, a collection of sensor data from the testbed in the Intel Berkeley Research Lab was investigated. The sample collected contains data generated by three sensors: (i) Sensor 4 - positioned in the center of the room, (ii) Sensor 16 - positioned in the upper corner, and (iii) Sensor 26 - positioned on the opposite corner to the sensor 16. Figure 31 shows the position of each sensor and highlighted in green are sensors 4, 16 and 26. These three sensor nodes collected measurements of temperature, humidity and light intensity.

All calculations began with data collected from Sensor 4, where two characteristics were noticed, (i) data is not ordered in a chronological way, (ii) there are several evident outliers. For example, temperature exceeding 30°C or below 10°C, and humidity inferior of 5%. Variables related to temperature, humidity and light are mixed in the same .csv file.

From a computational perspective, the methodology used in this study has sequential steps as follows:

Figure 31 – The position of the sensors 4, 16 and 26 are highlighted in green.



Source: Intel (2016)

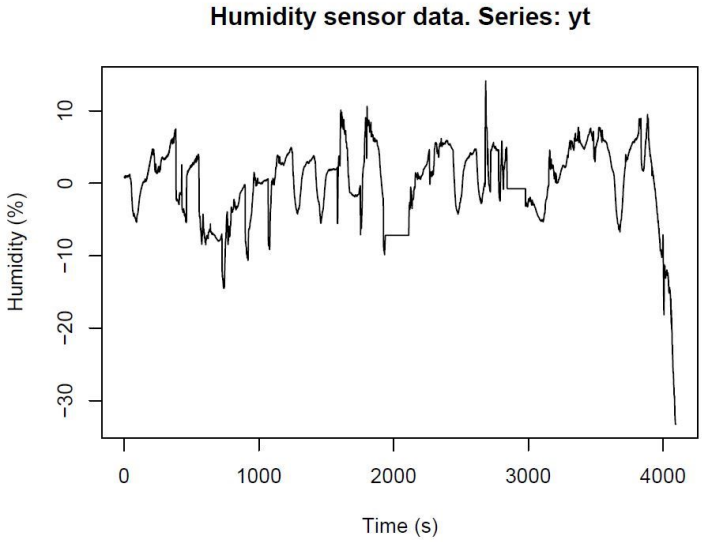
1. Our method chooses an initial sensor (Sensor 4);
2. Data about humidity, temperature and light are read;
3. The collection must be organized in equidistant data points; Data must be grouped in identical time intervals called Time Bins;
4. The total number of bins must be set by the user. In this research, we use 4096 $= 2^{12}$; twelve levels allow good precision in wavelet decomposition and good data compression;
5. The content of each bin is the mean value. However, it may happen to have bin without any data. In that case, the value of the previous bin is introduced because in this context the values change very slowly.

Figure 32 shows the result of the aggregation procedure for the Sensor 4, humidity data set, with its mean value removed.

We performed an exploratory data analysis of the signals to verify if they have the adequate characteristics to undergo a wavelet based denoising technique. Figure 33 shows the smoothed periodogram where we can note that the spectral density power increases as the frequency approaches zero. This indicates two important characteristics: i) the energy of the signal is concentrated in the lower frequencies; ii) the signal presents strong Long Range Dependence (LIMA; AMAZONAS, 2013). The first characteristic suggests that the signal may be cleaned by the wavelet-based denoising technique proposed in this work. The second characteristic offers guidance to model the signal for forecasting purposes.

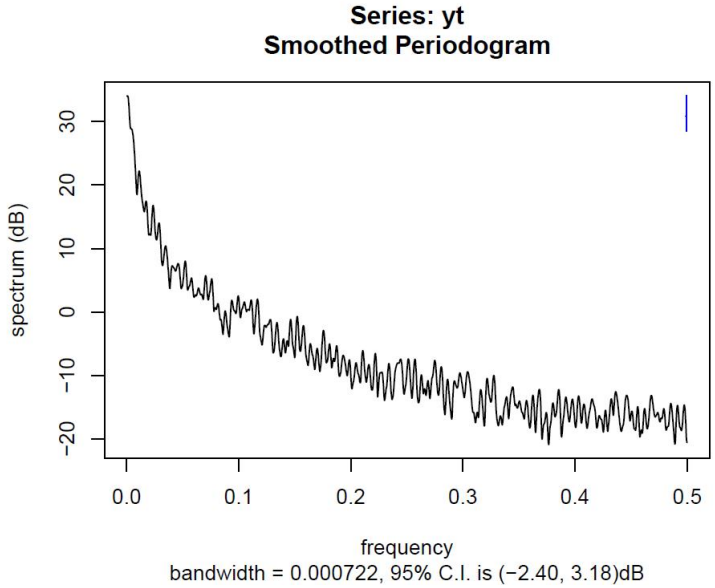
Figure 34 shows the vector of the DWT concatenated coefficients. The largest coefficients are concentrated at the end of the graph, i.e. at the low frequency region of

Figure 32 – Original signal received from the sensors aggregated in 4096 time bins.



Source: The author.

Figure 33 – Periodogram.

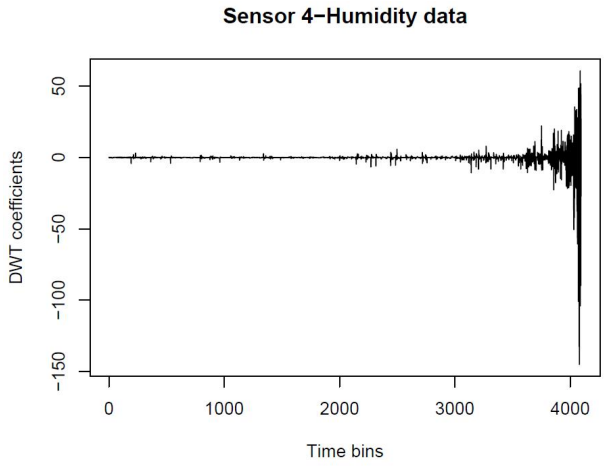


Source: The author.

the spectrum.

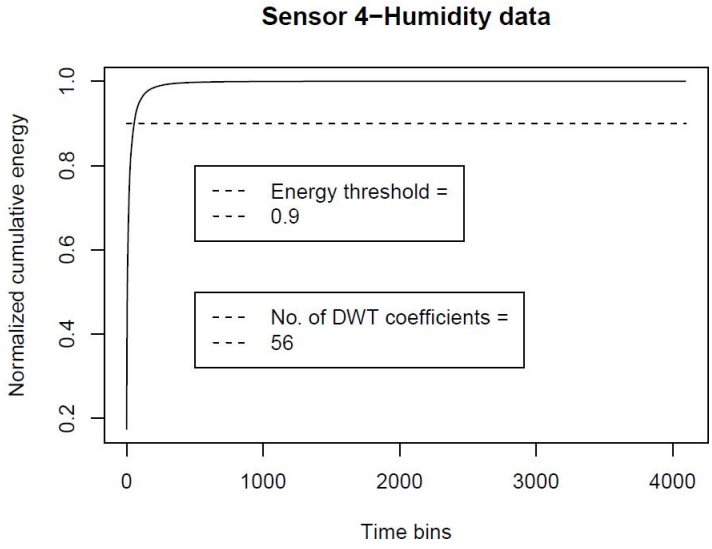
Figure 35 demonstrates the possibility of representing the signal using a small number of coefficients. In other words, only 56 wavelet coefficients are necessary to produce an approximate signal carrying 90% energy of the original signal.

Figure 34 – Vectors of coefficients concatenated.



Source: The author.

Figure 35 – Cumulative Energy.



Source: The author.

Table 1 shows the number of data points produced by the original signals. Table 2 shows the total number of wavelet coefficients necessary to produce an approximate signal with 90% energy of the original signal. In the worst-case (sensor 26 - luminosity signal) only 169 wavelet coefficients were needed from a signal that produced 61,520 data points.

The signal is reconstructed using only the necessary coefficients. The remained wavelet coefficients were set to zero as they carry more noise than useful information.

Figure 36 shows three graphs: (i) the top graph is the original signal from the humidity Sensor 4; (ii) the middle graph is the signal reproduced by all wavelet coefficients.

Table 1 – Original Data Set

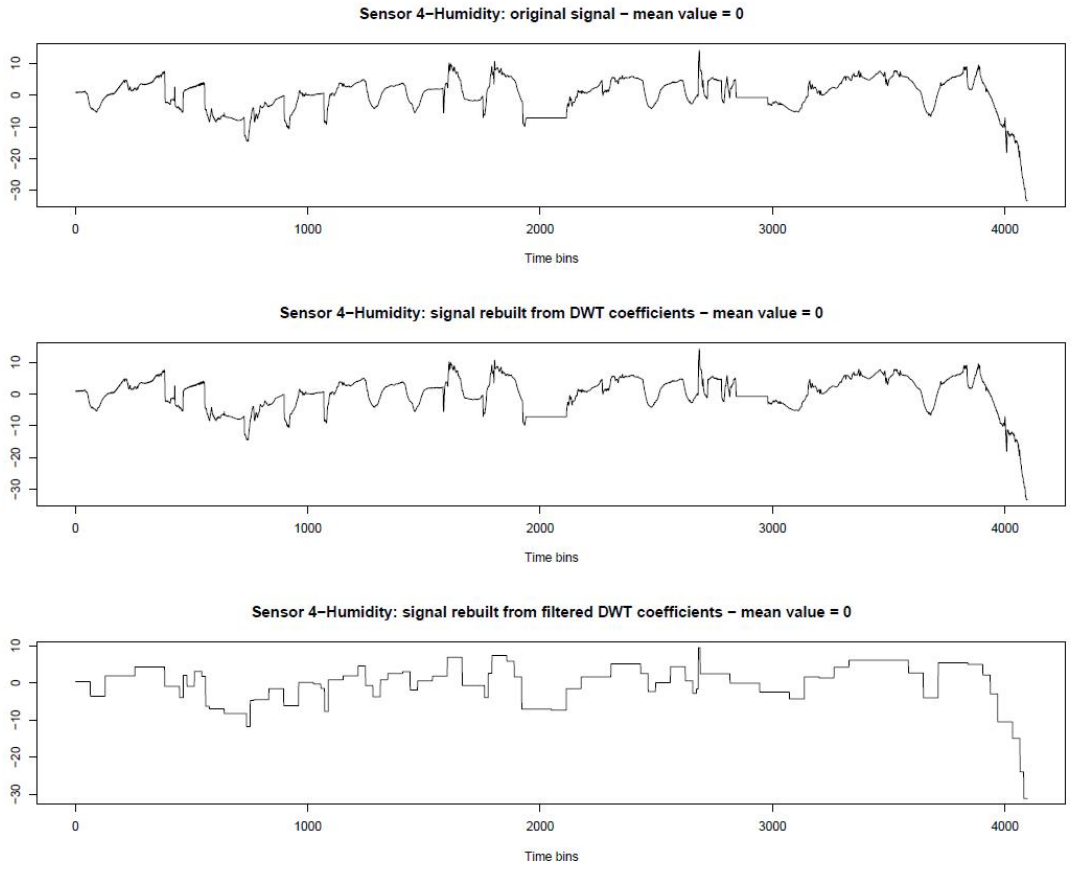
Sensor	Temperature	Humidity	Luminosity
4	32.966	34.604	43.790
16	26.109	26.292	34.582
26	41.778	47.318	61.520

Table 2 – DWT coefficients necessary to reconstruct the signal

Sensor	Temperature	Humidity	Luminosity
4	85 (0.258%)	56 (0.162%)	137 (0.313%)
16	78 (0.299%)	55 (0.209%)	128 (0.370%)
26	94 (0.225%)	75 (0.159%)	169 (0.275%)

This graph is identical the original graph (i) proving that the wavelet direct and inverse transforms were correctly implemented; (iii) the botton graph is the signal reproduced from the 56 largest wavelet coefficients. In summary, the approximate signal has a high degree of similarity when compared with the original signal.

Figure 36 – The approximate signal has a high degree of similarity when compared with the original signal.



Source: The author.

In agreement with this experiment, we found the same result for all sensors (in this case Sensor 16 and Sensor 26). It demonstrates that the wavelet denoising procedure drastically reduces the number of data points to be stored, transmitted and processed.

4.2.2 Application

Our analysis goes further than theory. This section aims to show that the filtered signal is helpful for developing applications. In other words, it is useful not only for reduction in the numbers of coefficient wavelets, but also to diminish the error in applications.

In spite of the good results obtained by the wavelet based denoise technique, one may argue that the approximate signal is a distorted version of the original signal and it could impair the development of applications. For this reason, an application was developed to show that the new denoised signal not only is useful, but also can improve the performance of applications. The sample contains data generated by the same three sensors: (i) Sensor 4 - positioned in the center of the room, (ii) Sensor 16 - positioned in the upper corner, and (iii) Sensor 26 - positioned on the opposite corner to the Sensor 16.

Comparing the graphs in the Figure 37, it is possible to see that Sensor 4 and Sensor 16 present different measurements of temperature. For this reason, the application aims to reduce this difference by making Sensor 4 control the temperature of Sensor 16. To put it more simply, the temperature of Sensor 4 was chosen as a reference and the temperature of Sensor 16 will be controlled by Sensor 4.

In this experiment, the application makes a temperature adjustment in the next instant (in Sensor 16), based on the temperature of the current instant measured at Sensor 4. That is, the application calculates the new temperature T_2 (in Sensor 16) based on the current temperature T_1 of Sensor 4 and based on the error that exists between them. The first experiment was done using the original signal and the second experiment was done using the new filtered signal (denoised by the DWT technique). The next statements are the logical steps followed by the application:

(i) Data from Sensor 4 and Sensor 16 were read;

(ii) The initial time and final time were set;

(iii) The data were acquired inside of the range of interest;

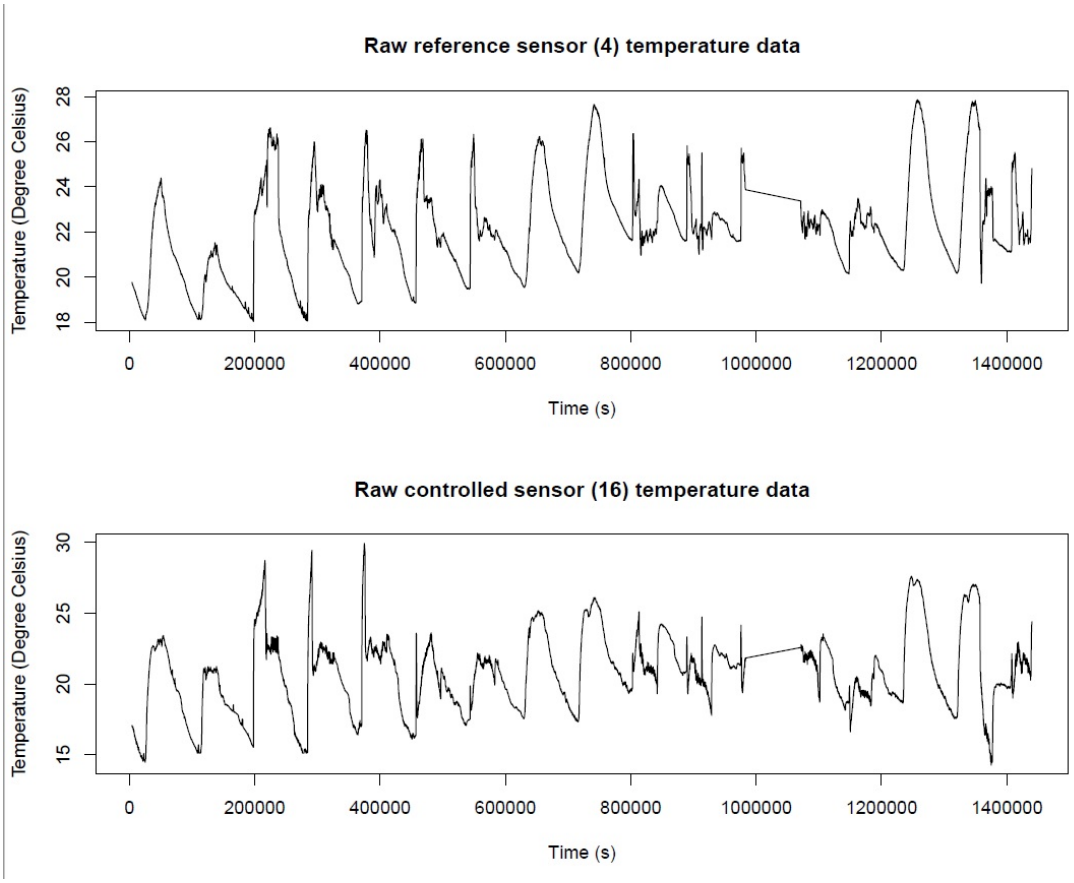
(iv) The data were aggregated in time bins (n = 4096);

(v) Reference sensor: $T_1(i)$, $1 \leq i \leq 4096$;

(vi) Controlled sensor: $T_2(i)$, $1 \leq i \leq 4096$;

(vii) The new temperature in the controlled sensor: new$T_2(i)$;

Figure 37 – Sensor 4 and Sensor 16 have different values of temperature.



Source: The author.

(viii) $\text{newT}_2(1) = \text{T}_2(1)$;

(ix) $\Delta\text{T(i)} = \text{newT}_2\text{(i)} - \text{T}_1\text{(i)}$;

(x) $\text{newT}_2\text{(i+1)} = \text{T}_2\text{(i+1)} - \text{k}_T.\Delta\text{T(i)}$;

(xi) Where $\text{k}_T \in \{0.1, 0.2, 0.3,....0.9\}$

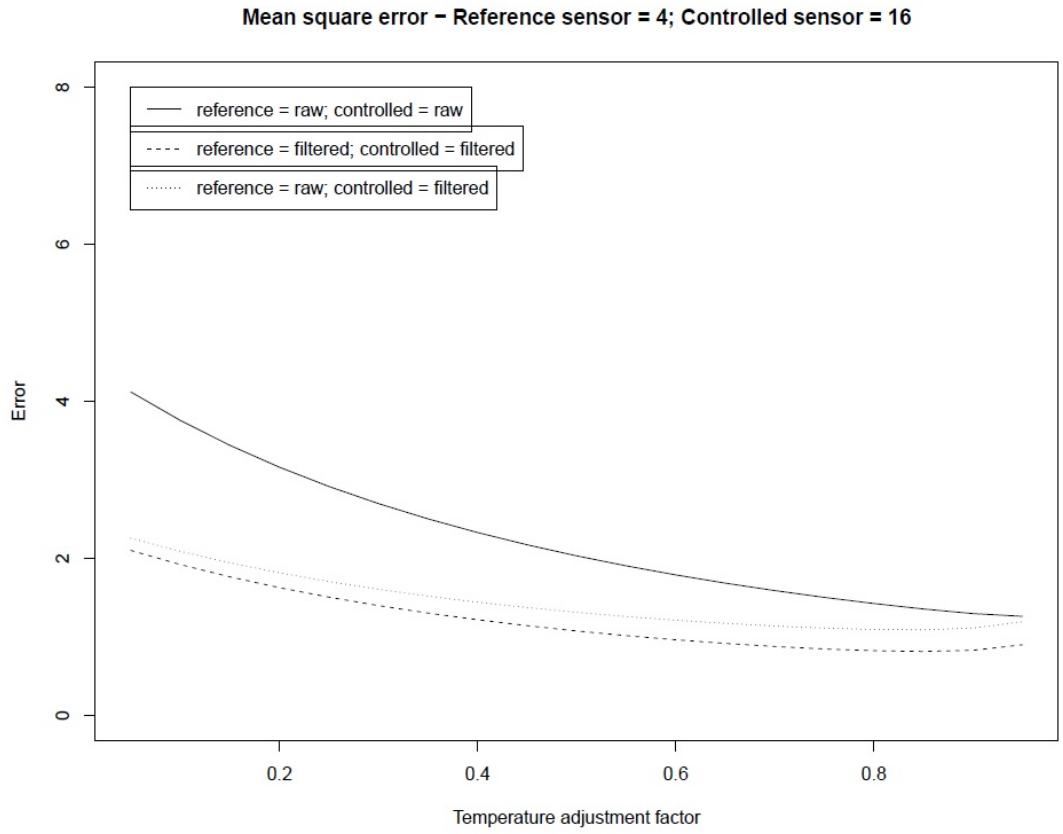
The method was assessed according to the following statement:

$\text{Error} = \sum_{i=1}^{n} (\text{newT}_2\text{(i)} - \text{T}_1\text{(i)})^2$

The graphs exhibited in Figure 38 show a comparison between three curves. These curves describe the performance of the application by means of three error curves:

(i) solid line: $\text{Error} = \sum_{i=1}^{n} (\text{newT}_{2,r}\text{(i)} - \text{T}_{1,r}\text{(i)})^2$, in which $\text{T}_{1,r}\text{(i)}$ is the original raw data and $\text{newT}_2, r\text{(i)}$ is obtained from the raw $\text{T}_2\text{(i)}$ data. This curve is taken as the reference because the denoising procedure has not been applied.

Figure 38 – Sensor 4 & Sensor 16. A comparison between the error produced by the orginal signal and the error produced by the denoised signal.



Source: The author.

(ii) dotted line: Error = $\sum_{i=1}^{n}$ (newT$_{2,f}$(i) - T$_{1,r}$(i))^2, in which newT$_{2,f}$(i), was obtained from a filtered (denoised) version of T$_2$(i).

(iii) dashed line: Error = $\sum_{i=1}^{n}$ (newT$_{2,f}$(i) - T$_{1,f}$(i))^2, in which T$_{1,f}$(i)), a filtered (denoised) version of the original raw T$_{1,r}$(i), was utilised as reference.

Clearly, the performance of the application, in terms of the final mean square error, always improves with the use of denoised signals.

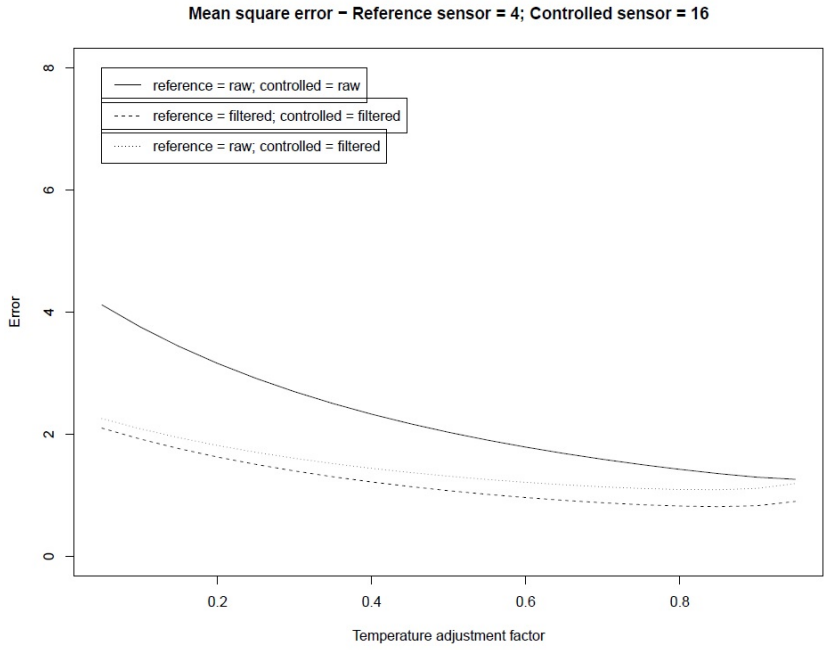
In summary, the error produced by the denoised signals are inferior than the error produced by the original signals. Therefore, the main advantage of applying the DWT technique is that applications can perform significantly better if this procedure is applied in advance.

In addition, the DWT-based technique has been applied to Sensor 16 and Sensor 26 where the same result was obtained. The experiment followed the schema:

- Figure 10: Sensor 4 controlling Sensor 26;

- Figure 11: Sensor 16 controlling Sensor 26.

Figure 39 – Sensor 4 & Sensor 16. A comparison between the error produced by the orginal signal and the error produced by the denoised signal.



Source: The author.

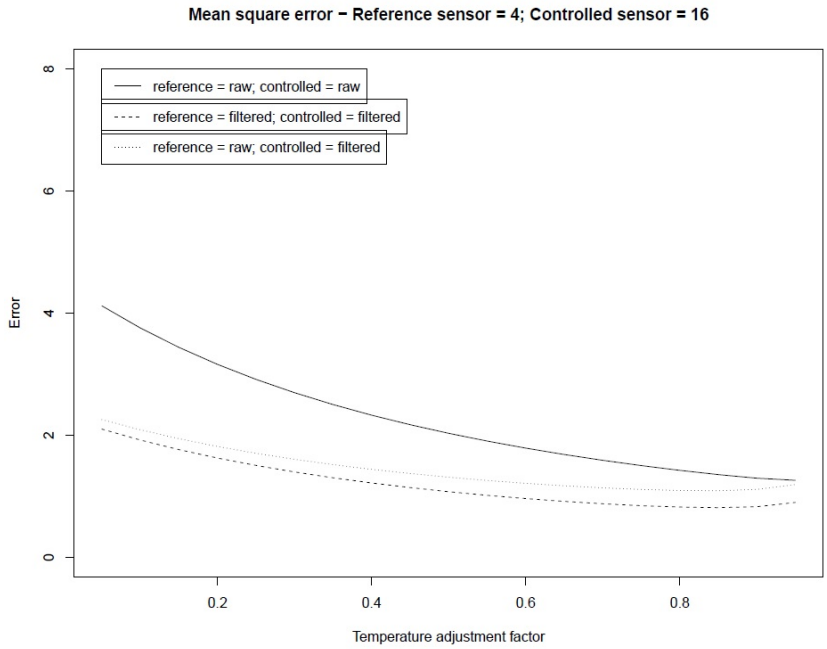
Comparing the graphs exhibited in Figure 39 and Figure 40, it is possible to conclude that the results obtained are in excellent agreement with Figure 38. In other words, in both cases the mean square error from denoised signals are inferior to that found in orignal signals. In summary, in all cases the denoised signals produced less errors.

4.2.3 Final discussion

In agreement with earlier studies, we found that noise is a non-essential signal that invades not only the Wireless Sensor Network (WSN), but also the Internet leading to transmission errors.

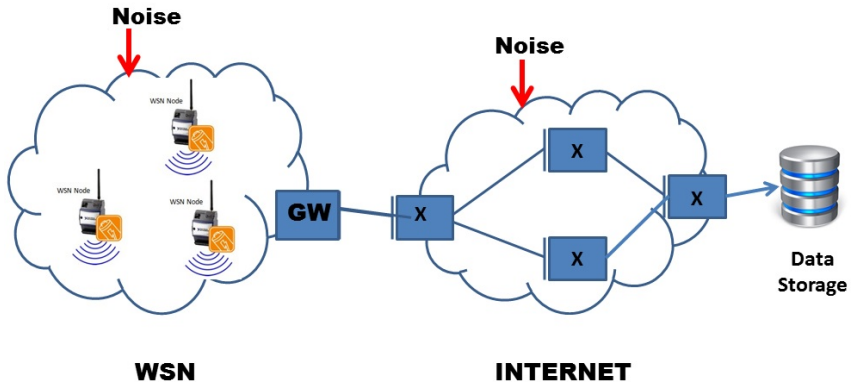
The experiemental evaluation presented in the previous sections, considered only the data set available in the Intel Research Lab, i.e, in the data storage shown in Figure Figure 41. Our recent investigation Nogueira e Amazonas (2012) demonstrates that, if we have access to the signal at the transmission side, then a DWT-based coding technique can be applied which improves the SNR (at the receiver side) by at least 3 dB. How to apply this DWT-based technique in the gateway (GW in Figure 41) will be subject of further investigation.

Figure 40 – Sensor 4 & Sensor 16. A comparison between the error produced by the orginal signal and the error produced by the denoised signal.



Source: The author.

Figure 41 – Noise can invade both the Wireless Sensor Network and the Internet.



Source: The author.

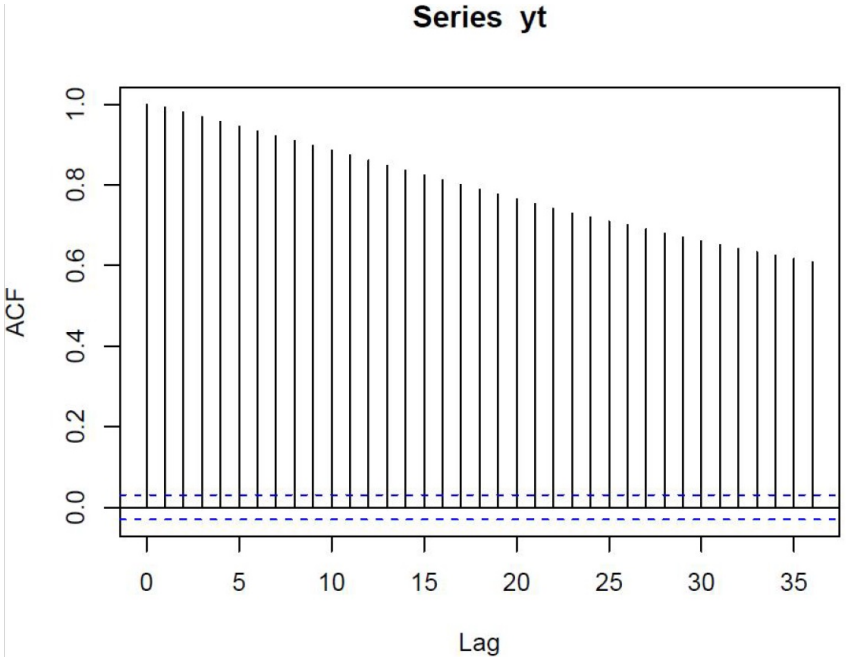
4.2.4 Predictions based on exploratory data analysis

Each measurement performed by the sensor produces a particular type of signal. When a sensor signal is cleared of noise, it is easier to predict because all coefficients that do not contribute have been eliminated. Therefore, it is possible to know the specific signal of each particular measurement. By knowing the signal is possible to assign a distinct semantic value for each one. Thus, when a new signal is produced by the sensor, it can be

compared with other signals and it can automatically receive the same semantic value of the more similar signal.

To make predictions using a sensor signal it is necessary to make some basic analysis utilising Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF), Variance and Periodogram were adopted. Although these techniques are very basic, they are essential for the analysis.

Figure 42 – Sensor 4 - Humidity Data - Autocorrelation function.



Source: The author.

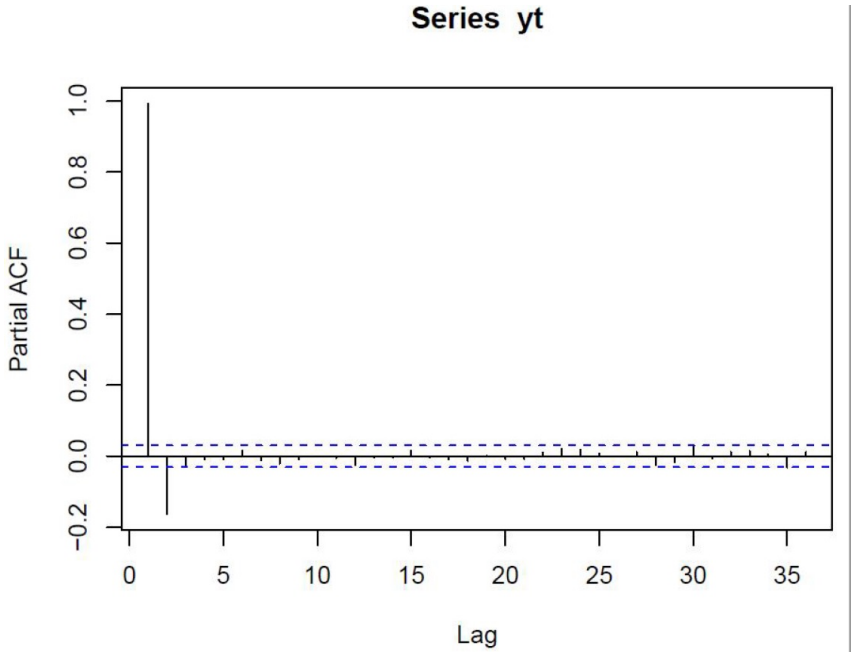
Figure 42, shows the graph of ACF, which is useful for measuring the degree of dependence between the values of a time series. The slow decay of the ACF in relation to the Lag indicates the presence of long-range dependence.

Figure 43 shows the PACF graph. The rapid decay of the PACF in relation to the Lag indicates the presence of short-range dependence, which can be modeled by a process called autoregressive (AR) with the lower order.

Figure 44 shows the Periodogram, where the rapid rise of the spectrum (dB), when the g tends to zero, proves the presence of Long Range Dependence (LRD).

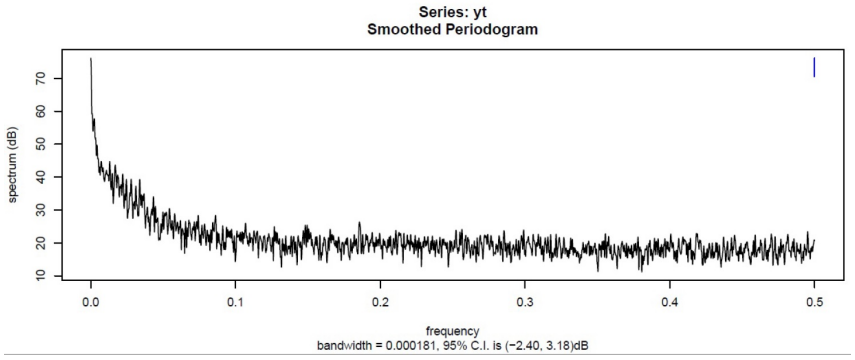
In a time series, the mean indicates the location of the gravity center. The variance measures the variability around its mean. At the top, the graph plotted in Figure 45 shows the presence of periodicity. The rapid raise of the line plotted proves the presence of long-term phenomenon.

Figure 43 – Sensor 4 - Humidity Data - Partial Autocorrelation function.



Source: The author.

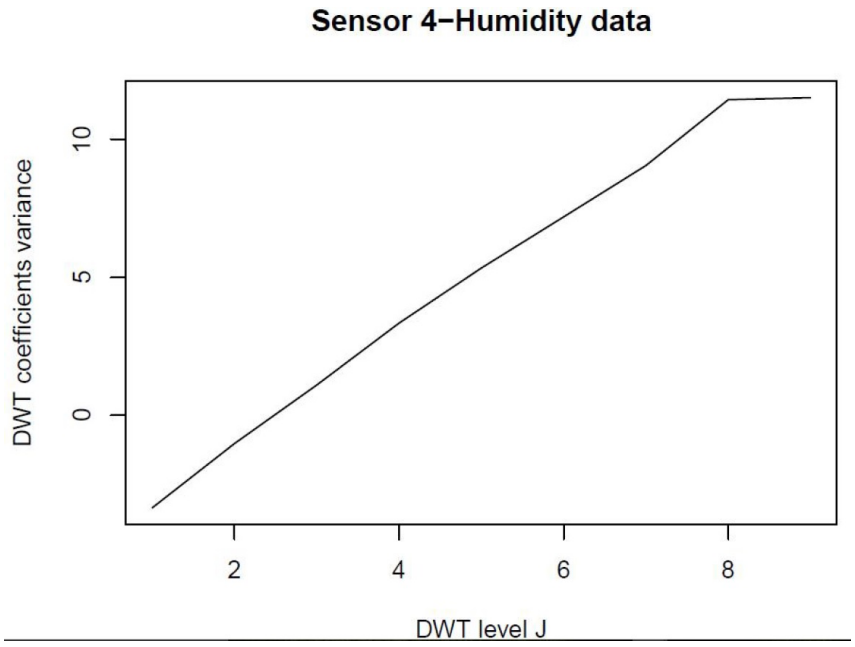
Figure 44 – Periodogram.



Source: The author.

Careful analysis of these graphs allows us to conclude that the data produced by this sensor signal can be modeled by a model type Truncated Auto Regressive Fractionally Integrated and Moving Average (TARFIMA) of a low-order. In brief, the exploratory data analysis can show the mathematical structure behind the time series indicating the best way to model the signal and consequently to develop forecasting tools.

Figure 45 – Variance.



Source: The author.

4.3 Web Data: Question Answering (QA) System

A good Question Answering System should be capable of providing correct answers to unambiguous questions asked by an individual in their natural language. This technique is useful to remove unneeded information quickly (VARGAS-VERA; LYTRAS, 2010). In comparison, keyword-based search engines, such as Yahoo and Google, present to the user a list of documents distributed in various web pages, while a QA system aspires to understand the questions in natural language, deduce the exact meaning and attempts to provide a precise and short answer (KONYS, 2015).

The QA system described in this section was developed by around 20 researchers and students during the Web Intelligence Summer School - WISS 2015 in Saint Etienne, France[1]. Participants were split into groups and each group developed a component of the QA system following the scheme presented in Figure 23. The practical section focused on:

- Natural Language Processing (NLP) Tools - (Syntactic Parsing, Semantic Role Labeling (SRL), Named Entity Recognition (NER), Relation Extraction);
- Semantic Matching (WordNet, Distributional Models);
- Semantic Web / Linked Data (Entity Linking. SPARQL, a query language for RDF, according to W3C [2]).

[1] WISS 2015 <https://wiss.univ-st-etienne.fr/>
[2] W3C define SPARQL:<https://www.w3.org/TR/rdf-sparql-query/>

These were the first tasks for each group:

1. Entity Search: Index the DBpedia graph using Lucene.
2. User Interface UI and QA Pipeline: Build the initial pipeline and the stubs for the components of the QA system.
3. Question Analysis: Using rules and regular expressions over POS Tags. Detect the lexical answer type of the example questions. Segment the question into a set of candidate terms. Use Stanford CoreNLP or NLTK.
4. Graph Extraction: Using OpenIE, extract relations from the Wikipedia articles Barack Obama, Paris, Jupiter.
5. Query Execution / Answer Ranking and Generation: Build an interface for the public DBpedia SPARQL Endpoint. Build a simple answer verbalizer from the SPARQL result set to a more natural language format.
6. Query Generation: Based on entity candidates and Stanford dependencies or C-structures. Build a triple-like representation of the query.
7. Evaluation: Using the latest QALD version, build a tool to calculate precision, recall and f1-measure for the example queries.

After three days planning, researching, coding and integrating the modules, the groups presented a QA system capable of answering simple questions, like: “Who is Nathalie Portmann?” This system has a server that is connected with DBPEDIA and is able to retrieve longitude / latitude and URL for the photos. The final interface is presented in Figure 46.

Figure 46 – A simple QA system developed during WISS.

Source: The author.

5 Smart scenarios

This chapter presents three scenarios where this research can be applied [1]. First, it provides a description of the location in order to show that apprenticeship can be maximised throug smart technology. Second, it introduces a solution based on this research. Finally, it demonstrates the benefits to the user.

5.1 Botanical Garden of Rio de Janeiro digitally augmented

In the latest edition of the *"Smart City Expo World"* (2013), Rio de Janeiro won the top smart city honours. The city has a Centre of Operations, it integrates data from approximately 32 municipal/state agencies and services under one roof, stopping the problem of the previously segregated and siloed city departments. Each area of the city is covered by the Centre of Operations. It has more than 400 active cameras monitoring incidents to provide safety and security.

One of the most popular tourist destinations in Rio de Janeiro is the Botanical Garden, also known as JBRJ. It is almost 200 years old and has been open to the public since 1822. It is located near the Corcovado Montain (north of Ipanema and Leblon) in the South Zone, a famous and populated area in Brazil. JBRJ is one of the most valuable tropical botanical gardens and arboretums of the world, occupying about 137 hectares, where 54 hectares are cultived gardens and 83 hectares are tropical forest. This vast garden contains more than 6.000 different species of plants and trees spread over 24 rows, 40 sections and 40 garden beds. Finally, it is the habitat of diverse species of fish, tropical birds and mammals.

Research and education are important activities of the JBRJ. There are postgraduate programs, support for undergraduate research, laboratories, environmental education and public outreach. Since 1915, the journal Archivos has been disseminating scientific studies undertaken in the garden. In addition, the Institution also distributes cds, books, journals, and the library has an extensive collection of specialized books, theses and dissertations. There are two types of tours: thematic and student. The thematic tour offers the following options: trail of the arts, trail of the noble trees and an historical trail. The student tour offers environmental education permanently, it has monitored itineraries covering multidisciplinary themes. In conclusion, this immense garden has a huge potential to provide learning about the species in its natural habitat.

[1] The development of these scenarios is being sponsored by PIPE Fapesp Fase I <http://www.fapesp.br/pipe>

As Perez-Sanagustin et al. (2014) have noted, visits and city tours are part of higher and secondary education, but are considered "less formal" than education in the classroom because it takes place in informal settings. According to the authors, Augmented Reality (AR) can improve the process of learning in the original location by offering digitally augmented learning settings. In other words, Augmented Reality can provide specific information by superimposing "digital" layers of information over particular place or object. This implies that, the AR combined with IoT and other digital resources can transform the JBRJ in a Smart Territory (which is already located in a Smart City). Moreover, Sentiment Analysis can provide to the JBRJ an overview about the experience of the students, this implies that, the institution can improve their services. In summary, these technologies working together can promote the process of learning in an informal environment and improve the quality of the services offered by the JBRJ.

The JBRJ has already a mobile application for tourists which is useful to guide them among Points of Interest (POIs) and to provide interesting information. However, they do not offer a mobile application for their students, such application could propitiate to learn a phenomenon exactly in the location where it occurs. It is important to note that, schools of all levels visit the JBRJ, in other words, many students can have more access to the information. Under this circunstance, this application aims to provide support in collaboration with Environmental Education.

The institution has a vast amount of digital information that could be transformed in very short sentences condensing an explanation. Its main idea is to summarise definitions, offering on time the most important key words about a targeted subject. Figure 47 shows a possible example of this idea. Giovannella et al. (2013) state that using ubiquitous computing one could learn in a smart, direct and active way.

Thus, the aim of this contribution is to provide to the students an immersive perspective of the JBRJ by offering them the opportunity to discover, explore and share hidden data in the landscape using mobile devices. In addition, the application can provide visual scientific names of the species (more than 6.000 different species) and quick explanations about them.

5.1.1 Possible solutions

The scenario described require the integration of many environments: physical, digital and social. Figure 48 presents this idea.

The system must be designed to achieve the following tasks: collect stream sensor data; collect local weather forecast data; fuse, interpret data and alert students about an interesting phenomenon. Enable students to collect, store, retrieve and share information. Equip the system to identify the student, interpret the data collected and shared; relate

Figure 47 – Scientific name of the species and essential information about it.

Source: The author.

his current emotional status to the professor and suggest activities based on the mood.

5.1.2 Virtual Environment

To achieve this goal, it is necessary to develop a mobile application where an educational content can be displayed in the natural place. Giovannella et al. (2013) suggest an application to assist users during the three stages of a territorial experience: "before", "during", "after". Consequently, the application must be divided into modules offering the following information:

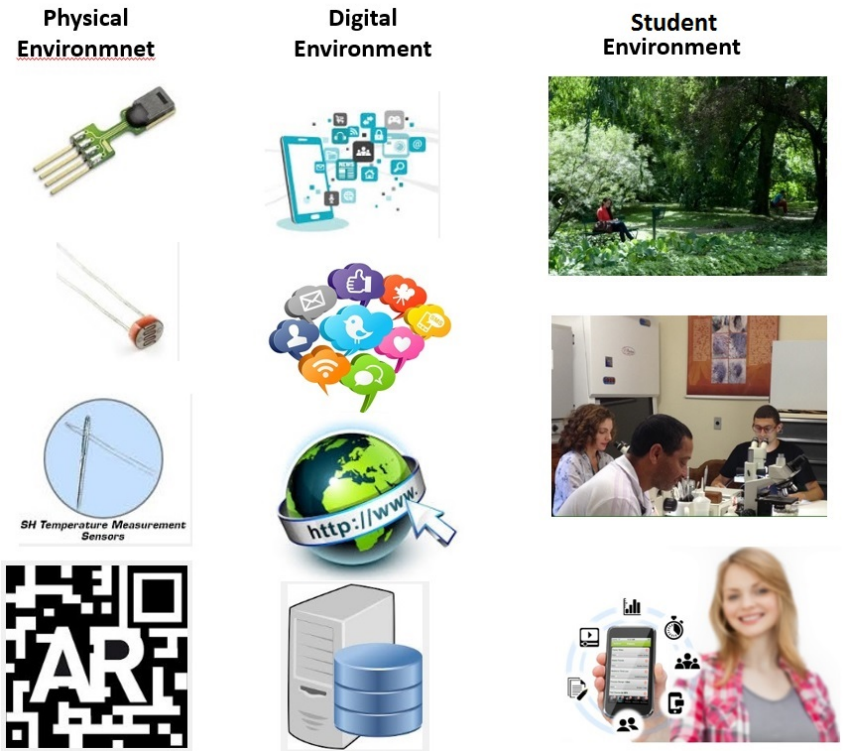
1. **Before**

    a) A summary of the research fields: botany, ecology, mycology, ethnobotanics, history.

    b) Information about laboratories: structural botany, seeds, algae growth, plant molecular biology, mycology;

    c) Weather forecast: is the day good or bad for exploratory activity?

Figure 48 – Data of three environments must be fused and comprehended.



Source: The author.

d) Sensors of temperature, humidity and light can monitor the POIs and indicate the most suitable occasion to manipulate plants; students also can use these sensor data to mesure how these factors are influencing the habitat.

e) Paths and points of interest - POIs; (different of touristic paths, these paths must be divided in accordance with specific research fields).

2. **During**

a) The Wikitude[2] augmented reality navigation system is the first pedestrian/car navigation system that integrates an augmented reality display and eliminates the need for a map. It can guide the student to the required place.

b) Mobile devices are used to read "markers" attached to or near each landmark. When a marker is detected AR presents summarized explanations about the species; Pokric, Krco e Pokric (2014) proposed a marker in the form of QR code.

c) As suggested by Bellini et al. (2013) students can vote "I like it" as a signal that they achieved an activity on campus and understand the explanation provided. Dislike option could suggest lack of understanding or doubts.

[2] Available on <http://www.wikitude.com>

Figure 49 – Wikitude is a type of navigation based on Augmented Reality.

Source: http://www.wikitude.com.

d) The app provides easy access for saving annotations, images and videos;

e) Students can contribute to the maintenance of the JBRJ, sending information and images about the natural conditions of the plants at certain geo-located point. Sensors can provide the conditions of the plants and give more bio-information to students and researchers.

3. **After**

a) Students can easily retrieve their notes, images and videos;

b) The app can offer an icon to redirect them to the mLMS, where they can answer questions or perform activities.

c) The student can also be redirected to the social networks in order to leave comments about the activities and share experiences. It is a common practice in various famous places, like the Red Cross Museum in Geneve-Switzerland, as shown in Figure 50. A simple *hashtag* like #sjbrj could be suggested, it facilitates the identification of the comments during the sentiment analysis.

d) Following the feedback provided by students, tutors can constantly improve the online content.

5.1.3 Physical Environment

Concepts of IoT-A and Semantic Web

The platform must provide an alignment between different environments in order to obtain reasoning and interpretation of data. The physical environment is a domain to be modelled. Ontologies have the ability to model content in a machine-comprehensible way

Figure 50 – The Red Cross Museum suggests a hashtag and offers interaction through Instagram.

Source: The author.

giving an exploitable meaning to the data. Therefore, each resource in the physical layer (e.g. sensor, actuator, processor, storage, tag) can be modelled using ontologies (composed of properties, classes and restrictions). IoT-A created a diagram to describe a resource. Figure 4.4 shows this model adapted from IoT-A project D2.1.

5.1.4 Citizen data: Understanding data in learning and e-learning contexts.

Khan et al. (2010) emphasize that it is necessary to diagnose the comprehension style of each student and their affective state because these factors influence learning. In addition, Rodriguez, Ortigosa e Carro (2012) indicate that affection and emotional aspects,

among other relevant elements to consider, affect student interest and the result of the learning process. According to them, in learning environments it is possible to measure feelings to better customize the learning process. In the same way, Ortigosa, Martin e Carro (2014) state that by identifying and managing information about the sentiments of the students for certain circumstances can help their potential needs at that point. In terms of e-Learning environment, the authors found several sets of basic emotions that they could consider and analyse:

- Joy, anger, disgust, fear sadness, surprise (STRAPPARAVA; MIHALCEA, 2008);
- Intellectually, empathy, spirituality (ANALYSER, 2014);
- Neutral, happiness, sadness, fear, anger, disgust, surprise Fulton (2014);
- Positive and negative (ANALYSIS, 2014);
- Joy, anger, fear, sadness (ZINCK; NEWEN, 2008);

The authors chose the last set of basic emotions because they are connected with "Fundamental Challenges" such as danger (leading to fear), separation from positive circumstances, along with insufficient self-efficiency (causing sadness), disappointment of expectancies and registration of inhibitions (provoking anger) or social acceptance and self-efficiency (generating joy).

Many work have been done in order to include Sentiment Analyses in e-Learning environments. Khan et al. (2010) included in a LMS the "Affective states and Learning Style" in order to identify learning styles and affective states and offer a personalized learning that combines the apprenticeship with the personal characteristics. Mao e Li (2009) included a Virtual and Affective Agent Tutor, called "Alice" which is aware of the sentiment of the student, while they are using the e-Learning environment, the system analyses their facial expression, speech and text, in order to adapt itself and provide personal learning in a conversational platform.

Not only affective tutor and emotional modules have been improving e-learning systems, but also text analysers like Chat-SEE (BUENO JUAN ANTONIO ROJO, 2011), a Semantic Emotional Evaluator for Chats. This tool, automatically evaluates the sentiments in online dialogues over a period of time. Furthermore, Rodriguez, Ortigosa e Carro (2012) developed a module to detect emotions from texts in order to use this information in an educational framework. The authors emphasize that associating emotions to these systems expand their possibilities, improving recommendations of activities according to the student sentiments at that time. In the same way, the system provides emotion-based content adaptation. Finally, the system is able to offer "motivation tasks" when the motivation of the student is low.

Ortigosa, Martin e Carro (2014) developed a tool to achieve sentiment analysis in Facebook aiming to improve e-learning systems. The first objective of the system is to recognize the polarity of the sentiment (positive, negative or neutral) in order to recommend activities appropriate to the current mood of the user. The second objective is to reveal critical emotional changes. This tool is called SentBuk, it captures messages, comments and likes of the user, shows the correspondent polarity and finally exhibits the sentiment of the user. Moreover, SentBuk presents the emotion of the friends, user categorization (according to their texts), statistics and other features.

5.1.5 Possible benefits of the system

The following benefits are expected:

1. Sensors collect real-time data at certain locations and provide intelligible information about the ecosystem;
2. Geographical data can provide the location of specific POI and conduct a student to it;
3. Sensors and geographical data help students to monitor natural events, send alerts about damaged or dead species specifying the location;
4. Tracking damaged or dead species facilitates the maintenance;
5. Students can collect, share and exchange information about the species in its natural habitat;
6. The system and tutors can track the experience of the students, their sentiments in that circumstances and recommend activities according their mood;
7. The recognition of the sentiment and tasks in accordance with current mood might improve the experience and apprenticeship of the student as state Ortigosa, Martin e Carro (2014).

5.1.6 Research & System Validation

The purpose of this activity is to verify if the system satisfies student needs and its intended use. Therefore, the following actions will be performed:

- First, it is important to carry out a Qualitative Research to understand what reactions students have to the new system. In most cases, this form of research is very useful to learn more about student preferences, opinions and reactions.

- Second, the validation process must include data collection during group discussion among students and personal interviews with tutors. In this case, data analysis will be non-statistical.
- Third, based on the initial understanding about student opinions Quantitative Research will be observed.
- Finally, the Quantitative Research can provide an understanding of what students think from a numerical and statistical point of view.

5.2 Using wearable devices to measure local happiness.

Data is an essential component in the process of acquiring useful information. In general, a person carefully examines a set of aggregated data before making an important decision. In fact, making decisions based on several sources of information can be very useful (EVANS, 2011) (SHULL, 2013) (DEVLIN; ROGERS; MYERS, 2012). Nowadays, a new flood of data are emerging from wearable devices, which are progressively being adopted to monitor, collect and send data about many events occurring in real time, from location to heart beat and other health metrics. The varied number of devices will produce different data types, each one describing specific details from which it will be possible necessary to extract information. However, very little research has been done on the aggregation of these data types. Consequently, this study investigates the fusion of wearable data. Here, the main hypothesis is the following: when data coming from wearables is fused with different data sources, new perspectives can arise, and it might be useful to the individual (RYOKAI et al., 2015). The next section indicates one possible scenario[3] [4] where fused data might be useful.

5.2.1 The World Happiness Report launched by United Nations

Since 2012, the United Nations - UN annually publishes the World Happiness Report (HELLIWELL; SACHS, 2016), the last edition was published in 2015[5]. In this report, the UN brings comparable data, analysis and the science of happiness to a wide global audience. UN highly recommends that countries start to measure the happiness of their population and use this to lead their public policies. According to them, politicians decisions could be based on better information and should consider scientific elements. Many countries, like Germany, the United Kingdom, South Korea, United Arab Emirates, etc. are quoted for their initiatives of making policies that can increase the levels of happiness. The report points out that when people are or become happier, a multiplicity of positive benefits can be realised, especially at the individual level.

5.2.2 The suggestion of an application

This research proposes a platform to measure local happiness by fusing data from multiple sources (ZHANG; TJHUNG, 1999). National and local governments, public and private institutions and hospitals, among others, can use happiness data in their selection of policies that could enable people to live better lives (DEVLIN; ROGERS; MYERS, 2012). The World Happiness Report states that there are six keys variables contributing to this measurement: (i) GDP per capita; (ii) life expectancy; (iii) social support (or having

[3] This scenario was presented as a paper and has been added to the 2016 ISCE Proceedings.
[4] This project was presented to the Desafio IoT 2016 and classified among the seven best projects.
[5] WHR15 <http://worldhappiness.report/wp-content/uploads/sites/2/2015/04/WHR15.pdf>

someone to count on in times of trouble); (iv) freedom to make life choices; (v) generosity; and (vi) absence of corruption (or perceptions of corruption). Specifically, it highlights three main classes of measurements:

1. Life evaluations;
2. Positive emotional experiences (positive affect);
3. Negative emotional experiences (negative affect).

5.2.3 Data coming from Social Networks

Positive and negative emotional experiences can be extracted from social networks, such as Facebook, Twitter and Instagram (ORTIGOSA; MARTIN; CARRO, 2014). Several scientific experiments that combine machine learning algorithms and several well recognised techniques to estimate emotional experiences have been done, producing good results (GHAG; SHAH, 2013).

5.2.4 Data coming from wearable devices

As Healthy Life Expectancy is a key variable to measure happiness, we believe that wearables can provide an overview of individual health in order to inspect issues in terms of health. These new types of sensors (wearables) can constantly monitor people bringing useful data related to their health (DOHERTY; LEMIEUX; CANALLY, 2014). Some of the most well known wearable devices are: smart watch, smart scale, gadgets for activity tracking, and wireless blood pressure measurer. The desirable features are: continuous heart rate or wrist heart rate, step tracker, running physiological measurements, built-in Global Positioning System (GPS) tracker, and automatic sleep detector and a scale to measure the weight, bone mass, body fat, water, muscle, etc. Preferably, these gadgets will automatically sync wirelessly with the Local Area Network (LAN) or Personal Area Network (PAN) and other devices. Some recommendable devices to perform this experiment are: Garmin Fenix 3 HR - Smart Watch, Fitbit Surge Fitness Superwatch, Withings Wireless Blood pressure monitor for apple and android, Wireless smart scale BlueAnatomy and Skulpt chisel.

5.2.5 Data coming from web sources

Daily news online can demonstrate a panorama of the country, showing how much the levels of Corruption and Changes in the Quality of Governance is affecting the national happiness. According to UN, these two elements can affect the mood of the population significantly.

5.2.6 Data analysis and technical view

The fulfilment of these principles is leading to various design challenges. First, in regard to individual data, the application will allow connection to several social networks at the same time, find user location automatically (if not possible, the user can select it). The application will detect automatically a new post and identify the author, day / hour of the post, detect the topic and the mood of the user. Second, relating to Sensor Data, there are several devices to bring up-to-date individual data and provide an overview of Healthy Life Expectancy. Probably, users will connect various different personal wearables. Therefore, it will be mandatory to collect, transform, store, and retrieve data from a variety of gadgets representing many different operating systems, standards and protocols. Sensor data has many challenges. As can be expected, it needs to be individually analysed and integrated in real time in order to extract value promptly. Web Data allows more information from several websites be included. Information can be included, pop-up automatically, or be available as default, for example weather forecast, local traffic, important events in the city, etc.

5.2.7 Steps to connect the application

- Users will download the mobile application to their mobile devices.
- Users will provide identification through login and password.
- The application automatically recognizes location.
- Identify applicable (local) sponsor/s.
- Users can choose the feeds they want between the default option or the set menu.
- A platform to visualize all data coming from the population must be designed.
- Using this platform policy makers may have an overview of the local happiness.
- Through observation, history and intelligence, it can provide insights and predictions.

5.2.8 Benefits to the user

According to the UN World Happiness Report, the main foundation for good policy in a democracy should be to maximize the happiness of the citizens. UN affirms that nowadays policy is based on little more than subjective assessments/opinions and more scientific elements should be explored. Moreover, politicians should focus on existing evidence on what promotes happiness and metrics to monitor what actions are effective. For this reason, we propose a digital tool that can help policy makers have insights about

local happiness. Moreover, individuals can discover their own levels of happiness and compare their personal happiness with the local happiness, developing plans to improve it.

5.2.9 Wearables devices that will be tested in this scenario

UN recommends focusing on existing evidences on what causes happiness and what actions are effective.

Politicians should be aware of the impact of their actions and frequently ask: "Are they happy with their lives, are they satisfied?" As stated in the report, the right judge of each citizen life is that same citizen. UN recommends focusing on existing evidences on what causes happiness and what actions are effective. For this reason, this designed scenario propose a digital tool that can help policy makers have insights about local happiness and compare their personal happiness with the local happiness, developing plans to improve it.

5.3 Data fusion in intelligent transportation systems: using the citizen as a sensor

One of the main research themes at the moment is related to smart cities, which includes the implementation of the IoT paradigm in cities. According to Giffinger (2007), there are six characteristics of a smart city: smart economy, smart people, smart governance, smart mobility, smart environment, and smart living. The concept of Intelligent Transportation Systems (ITS) is part of the smart mobility scope, and it is defined by Iovanovici, Prodan e Vladutiu (2013) as systems that are composed by hardware, software and people that have as their main objective to monitor and react to the conditions of transportation networks. Well designed systems can affect the everyday life of citizens, positively impacting on their well being. Among the services of ITS, it is possible to observe: adaptive personalised maps and vehicle navigation, smart fleet management, traffic monitoring, congestion avoidance, road incident detection, among others (POSLAD et al., 2015).

All these scenarios described above shows that data and information can be generated at any moment, anywhere, by anyone or anything. The varied number of data sources will produce different data types, each one describing specific details from which it will be possibly necessary to extract information. As data proliferate and volumes grow, the extraction of relevant information can become very difficult, because a significant number of data of little importance emerges alongside the main information.

For these reasons, this work investigates the fusion of three data sources: data produced by citizens, web data and data from social networks, to improve the citizens' experience in public transportation system, in an ITS context. The following hypothesis is made here: the platform can improve transportation planning for the citizen, the transportation companies, and the city administration, leading to better decision making and optimization of the transportation system. Thus, here is proposed a mobile application that will collect and receive data that would contribute to the main stakeholders on their decision making processes in the long, medium and short term.

5.3.1 Requirements identification

Besides having the citizen in its center (instead of a company or a government agency), other important requirements that were identified were: ease of use, because most of the stakeholders are not experts in computer programming or data analysis; a platform of visualization with friendly interfaces to the major stakeholders (citizen, city council and transportation companies); the need to consider the information needed for both long, medium and short-term planning for all stakeholders; the need to impact as little as possible on the current habits of the citizen; and the possibility to incorporate

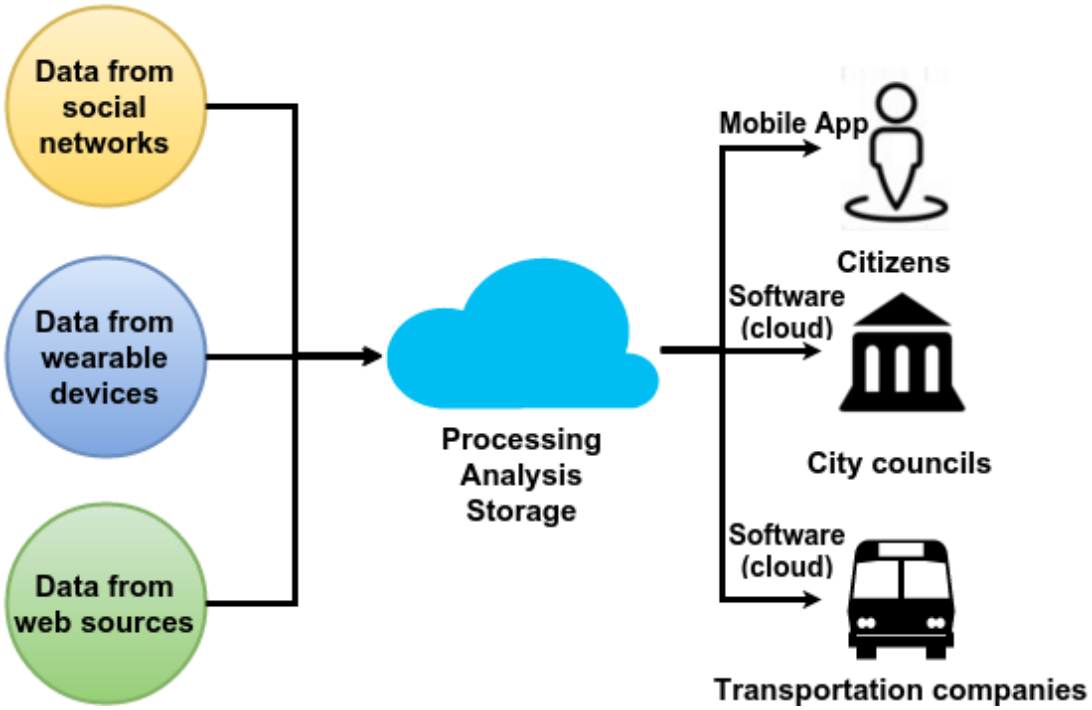
data from different sources: social networks, sensors devices, and web sources.

For these reasons, the proposed platform is divided in two different platforms: a mobile app that will be used by the citizen, and software in the cloud, that will be used by the city councils and the transportation companies. If needed, other stakeholders like NGOs and Government agencies could also have access to this platform.

5.3.2 Main characteristics of the proposed tool

The information provided to the users of the digital tool can be used in different categories of decision making: strategic (annual planning), tactical (monthly and weekly planning), and operational (daily planning). A summary of the decisions in each of these categories for the three main stakeholders are illustrated in Figure 51.

Figure 51 – A general overview of the idea.



Source: The author.

In the case of the citizen, the mobile app will provide information that will be useful for making better choices related to: choosing which type of transportation will be the main one used for different purposes (going to work, recreation and leisure activities); what routes the citizen should use on a daily basis, as part of a weekly planning (considering infrastructure maintenance or development); what alternative routes the citizen could use in case an unplanned event (such as a flooding or an accident) occurs; and what route the citizen could use for a route he or she has not used before, as in the case of sporadic trips. Nowadays, most of these decisions have to be made without historical data, and it

is difficult for users to access all the applications and websites that would allow them to gather the necessary information. So, the solution described here could provide him or her an easy-to-use interface with close to real time results, improving the user experience and, as result, the value added by the transportation system.

In the case of the city council the software in the cloud will provide information that will be used to decide: what infrastructural projects are more important for the city; how to best coordinate these projects (in a public, private, or partnership modes); how to design the transportation system, considering the historical changes in the routes' volume of use; which transportation companies to select and how to better evaluate them; how to monitor the transportation system on a daily basis; and how to invest the resources of the council to provide the cost effective solutions to improve the well-being of the citizens.

The last major stakeholder in this system is the transportation companies. They will use the same software provided to the city council (but with different interface and information access) to make better decisions related to: investment on new and existing equipment; how to best allocate the buses on different routes based on the volume of use (and how to change this over time); which routes should have more or more frequent buses, and at what times of the day; and what are the alternative routes that can be used in the case of unplanned events.

In the case of other stakeholders that may be interested in accessing the information provided by the platform, we propose a different interface that will also be hosted in the cloud, but that will give access to limited information. If more information is needed, they will be able to use it to request to the responsible entities.

The next subsection contains a description of the platform proposed, focusing on the mobile app that is being designed for the citizens.

5.3.3 Description of the platform proposed

The platform proposed will use the cloud to process, analyse and store data from social networks (mainly Twitter, Facebook and Instagram), smart devices (mobile phone and additional sensors that citizens can use to send data) and web sources (current digital news, weather forecasts, and official information transmitted via Internet by the city councils and transportation companies).

The information necessary for the users of the tool will then be provided via mobile app (for the citizens) and software in the cloud (for the city councils and transportation companies). The main aim of this design is to minimize the need to install additional software or to develop new habits for sending and receiving data by the users.

Table 3 contains the main data collected for the three stakeholders by the digital tool we propose, considering the communication technology and the sensors used.

Table 3 – Main data collected, sensors used, communication technology, and information provided for the main stakeholders in our proposal

	Citizen	City council	Transportation companies	Web sources
Data collected	Login information	Traffic information	Bus timetable	Weather forecast
	QR code reading	Info for the citizens	Estimated bus arrival times	News
	GPS reading (if no QR code)		Bus GPS position	Social media posts
	User reports		Additional bus company info	
Sensors used	Cellphone GPS	Field data collection by the traffic agency	Bus GPS	None
	Cellphone camera			None
Communication technology	3G/4G	3G/4G	3G/4G	Internet
		Internet	Internet	
Information provided	Bus arrival time (estimated)	Main routes/time of the day (estimated)	Number of readings/bus stop	None
	Alternative routes/buses		Number of users/bus line/time of the day (estimated)	
	Other relevant information	User reports		
	Recommendation of alternative routes for the day			

It is important to observe that our tool relies on the ownership of smartphones with access to 3G or 4G network by the citizens, what is already a reality in the city of São Paulo, Brazil. Another premise that was considered is that transportation companies and the city council will provide information related to the buses timetables and estimated bus arrival times, due to the fact that there are already commercial mobile applications that have access to these data. As Figure 52 shows a summary of the main idea.

The major aim in this paper is to improve the experience and wellbeing of the citizen, we focused our efforts on designing the mobile app that will be used by them. This is illustrated on Figure 53.

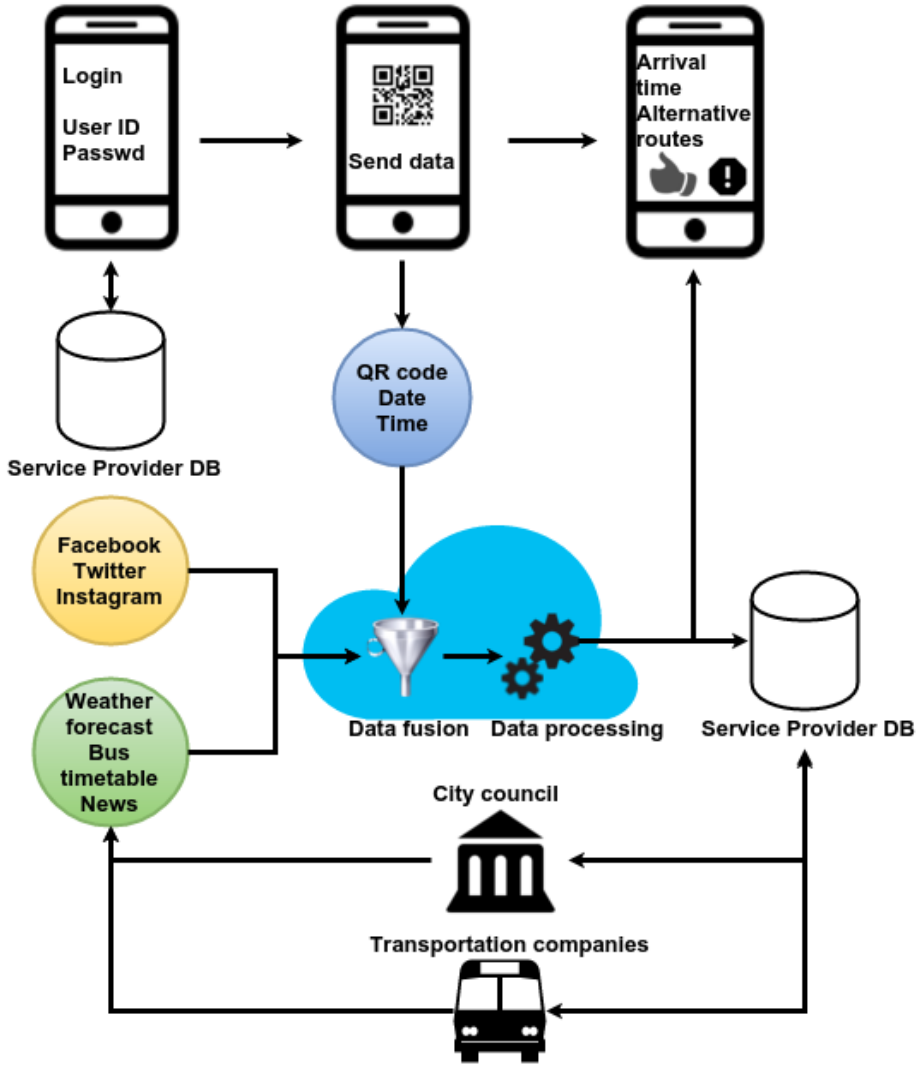
In the vicinity are Web Data, where detailed information from several websites can be included. Information can be included, pop-up automatically, or be available as default, for example weather forecast, local traffic reports, important events in the city, etc. In the following section, operation of the application is described.

First, the user will download the application at the appropriate application store. Then, he or she will create an account, with a user ID and password. They will be able to insert personal information, or just provide an email for password recovery processes. This data will be stored in the service provider database.

After logging in, the user will be asked to send data to identify the location of bus stop where they are located. After considering several alternatives, it was decided that the best option would be to attach QR codes on the bus stop or if these are not readable (or a mobile camera is not working), to allow the citizen to use the GPS embedded in their mobile phone to provide a location. As a last option, the user will be able to manually insert the address, and it will be provided as an autocomplete function via interaction with the Google Maps API.

The data containing the QR code and the current date and time will be sent to the cloud, in which a data fusion process occur, considering also the information obtained and pre-processed from the social media platforms (Facebook, Twitter and Instagram) and web sources (weather forecast, bus timetable and news). These data will then be processed, and the resulting information containing an estimation of arrival time and alternative routes (in case an unplanned event occurred, such as a flood) will be sent to the citizen.

Figure 52 – Design of the proposed digital tool.

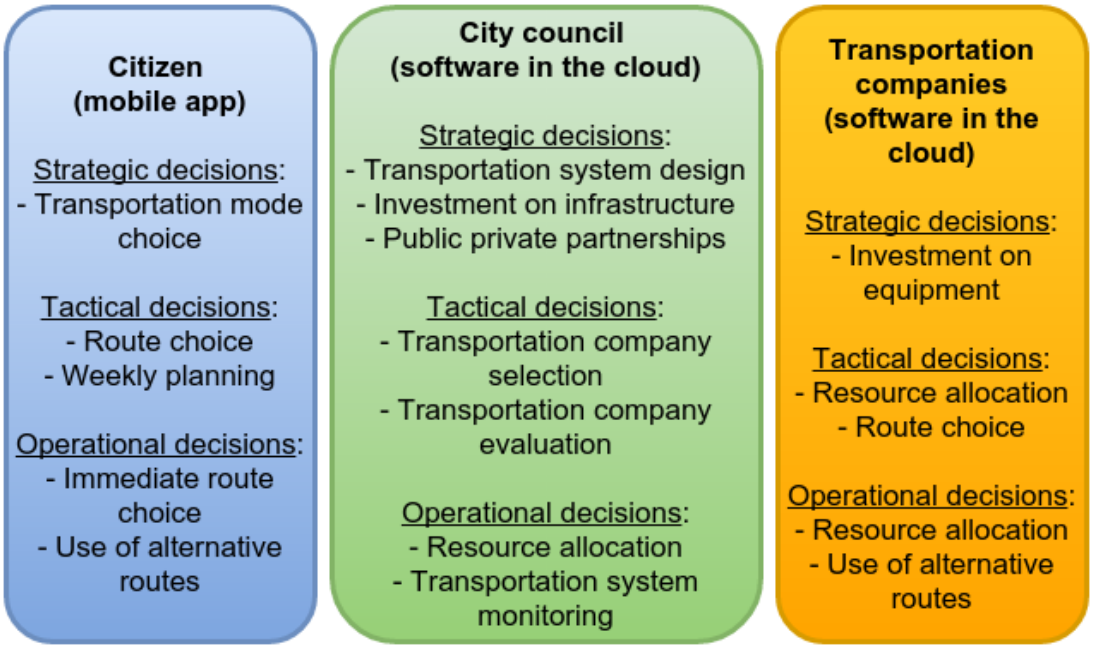


Source: The author.

The citizen can then indicate *"like"* for the answers, report an error, or provide suggestions for the city council, the service provider, or the transportation companies. All the information generated and data collected from the citizen will be uploaded to the service provider database.

Through observation, history and intelligence on the data and information generated, the design proposed can provide insights and predictions to address the strategic, tactical, and operational decisions of the main stakeholders identified. Because it focuses on the citizen instead of on other agents, it might be able to improve their wellbeing.

Figure 53 – Design of the proposed mobile app to the citizen.



Source: The author.

5.3.4 Limitations and future works

The main limitations observed during this project were: the lack of data available from both the government and the transportation companies related to the behavior of the citizens and their average waiting time for the different bus lines, and the difficulty to contact the city transportation management team. Future works are related to: development of an App (application) for the Android operating system; conduction of an experiment to test the validity of the application proposed; development of the software that will be used in the cloud by the transportation companies and city councils; and the elaboration of a system dynamics study to understand more deeply the factors that impact the behavior of the citizens towards shifting from private to public transportation.

6 Conclusion

Nowadays, not only are people producing data, but also companies, sensors, etc., are propagating data like never before in the past. Based on these informations, people take decisions, make deals, change routes, call, text, express opinions and sentiments, generating more data and information. For this reason, the aim of this thesis was to enhance the experience of the user (or individual) by providing a broad perspective about information that matters to him/her, which could result in improved ideas and better decisions.

The present study investigated three data sources, their characteristics and issues, the suitable techniques to deal with them and a scheme for integrating these data sources in a platform. This study demonstrates that each data source has specific challenges and details that must be separately analysed. Moreover, these sources present several important characteristics that must be identified and tackled in advance.

Initially, the data integration was conceived as part of this thesis. However, the process of integration of these three data sources has been a big challenge. It is difficult to assemble all elements, making each component of the system function properly as a whole, reach the design properties, specify interfaces, while performing the verification and validation of each part. As can be expected, these tasks have a big demand for different expertises, robust infrastructure of hardware and modern platforms for software development. For this reason, this part of the project is being sponsored.

This research has shown that when analysing individual data, around 700 profiles from Twitter were collected. These profiles were clustered in accordance with the words found in each profile. By clustering the profiles, it is possible to identify communities that share similar interests, which means that individuals belonging to one specific community could receive similar information. Other data sources must be investigated in the future. In Brazil, the most popular social network is Facebook. Therefore, in the near future this data source must be analysed and integrated into the platform. Other social networks such as Instagram and Pinterest are being suggested.

Our observations that sensor data faces many challenges, especially in terms of volume and noise, are not new. The main finding of this area was a very impressive technique that is capable of filtering a sensor signal, significantly reducing the size and the noise of the data. This research assesses the impact of the filtered signal when developing applications. The results show that it is useful to diminish the error in applications. Other gadgets could be analysed in order to discover if their data is really useful when integrated.

The last data source analysed was web data. Among several optimal options, a Question Answering System was the technology chosen for the platform. This is because

QAS is the most popular system at this moment, being adopted by large companies such as IBM, Apple, Microsoft, etc. The QAS presents various challenges, especially because questions are presented to the system in natural language. A simple prototype was developed, tested and it is ready for use. This prototype has a number of possible limitations, but simple questions can be posed, meaning that information can be available easily and quickly.

Several questions remain to be resolved, in particular how to integrate all these techniques in the proposed platform, create useful Apps, develop smart dashboards to managers and authorities, test the experience of the users, etc. These questions suggest that the next phase will also involve different expertise, robust hardware and databases, up-to-date software, etc. For this reason, monetary support was requested.

6.0.5 Contributions

The main contribution of this thesis consists in systematically introducing the user (Individual) in the center of the system. In other words, individual data is the core of the system. The other data sources are aggregated in accordance with the individual perspective. This user-centered approach can be adopted because the hardware and technology are mature and the focus can be directed to the user. In the same way, the platform proposed take into consideration the requirements and needs of the Individual, giving primary attention to their data when integrating related data.

Also, we proposed an innovative technique to improve the signal of a sensor, enabling a more trustworthy extraction of data. The proposed technique was based on Discrete Wavelet Transform which aims to clear the wavelet coefficients that contribute most to the noise than for the signal. The present research shows that it is possible to reduce signal noise using discrete wavelet transform, first decomposing the signal in levels; second, generating vectors of coefficients; third, choosing the most energetic coefficients to reconstruct a new signal using Haar as the mother wavelet. The graphic of the approximate reconstructed signal (with 90% energy using only 56 coefficients chosen from among 34,604) shows a signal that appears very similar to the original. We developed an application that shows that the approximate denoised signal not only is useful, but also improves the error performance

Results has shown that it is not effective for all signals because the signal must have favourable spectral characteristics, i.e., the information regarding the signal must be concentrated in small number of wavelets coefficients. In summary, DWT is a powerful tool for reducing the volume and noise of sensor data.

6.0.6 Future Work

Future research should focus on the fusion of data, taking into consideration correlated data while giving the major attention to the data produced by individuals. It will be possible by including data sources like Facebook, Instagram and Pinterest. Tests of accuracy, interfaces, security and privacy must be performed and bugs removed. It is also necessary to verify if the system satisfies individuals need and its intended use.

App for smartphones and visual analytics tools (or dashboards) must be developed and tested.

The recent suggestions about geo-tagged tweets are worth exploring, and we will do so in future work.

Also, we intend to extend the exploratory data analysis to model the sensor signal for forecasting purposes.

6.0.7 Final Considerations

People from all over the world aim to access information on varied topics, all the time. It means that accurate information must be available to anyone, anywhere, anytime. For this reason, the new systems must be prepared to deliver precise and specific information to one person, taking into consideration the topics that matter to him/her personally. Today this scenario is possible by collecting data from various sources, integrating them in a platform while using a user-centered approach. By putting the Individual at the heart of the system it is possible to enhance their experience.

References

AKAY, A.; DRAGOMIR, A.; ERLANDSSON, B. E. Network-based modeling and intelligent data mining of social media for improving care. *IEEE Journal of Biomedical and Health Informatics*, v. 19, n. 1, p. 210–218, Jan 2015. ISSN 2168-2194. Citado na página 27.

ALLEMANG, D.; HENDLER, J. A. Semantic web for the working ontologist - effective modeling in rdfs and owl, second edition. In: . [S.l.: s.n.], 2011. Citado na página 35.

ANALYSER, H. Emotional marketing value headline analyzer. In: . [s.n.], 2014. Disponível em: <http://www.aminstitute.com/headline/>. Citado na página 85.

ANALYSIS, S. S. Emote api. In: . [s.n.], 2014. Disponível em: <http://www.sas.com/en_us/software/analytics/sentiment-analysis.html>. Citado na página 85.

ASIN, D. G. A. *50 Sensor Applications for a Smarter World*. 2012. Disponível em: <http://www.libelium.com/top_50_iot_sensor_applications_ranking/>. Citado na página 29.

BARNAGHI, P.; SHETH, A.; HENSON, C. From data to actionable knowledge: Big data challenges in the web of things. *Intelligent Systems*, v. 28, n. 6, p. 6–11, Nov.-Dec. 2013. Disponível em: <http://ieeexplore.ieee.org>. Citado na página 19.

BARNAGHI, P. et al. Semantics for the internet of things: Early progress and back to the future. International Journal on Semantic Web and Information Systems, v. 8, p. 1–20, 2012. Citado 2 vezes nas páginas 31 and 45.

BELLINI, A. et al. Once upon a time: A proof of concept augmented reality collaborative mobile application to discover city heritage. In: *Signal-Image Technology Internet-Based Systems (SITIS), 2013 International Conference on*. [S.l.: s.n.], 2013. p. 358–363. Citado na página 82.

BENZING, A. et al. Multilevel predictions for the aggregation of data in global sensor networks. In: *Distributed Simulation and Real Time Applications (DS-RT), 2010 IEEE/ACM 14th International Symposium on*. [S.l.: s.n.], 2010. p. 169–178. ISSN 1550-6525. Citado na página 30.

BERKNER, K.; WELLS, R. O. Wavelet transforms and denoising algorithms. In: . [S.l.]: IEEE, 1998. v. 2, p. 1639–1643 vol.2. ISBN 0780351487, 9780780351486. Citado 4 vezes nas páginas 30, 46, 53, and 64.

BEYKIKHOSHK, A. et al. Data-mining twitter and the autism spectrum disorder: A pilot study. In: *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. [S.l.: s.n.], 2014. p. 349–356. Citado na página 26.

BHATTACHARYA, A.; MEKA, A.; SINGH, A. K. Mist: Distributed indexing and querying in sensor networks using statistical models. In: KOCH, C. et al. (Ed.). *VLDB*. [S.l.]: ACM, 2007. p. 854–865. ISBN 978-1-59593-649-3. Citado na página 31.

BHOIR, V.; POTEY, M. A. Question answering system: A heuristic approach. In: *Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on the*. [S.l.: s.n.], 2014. p. 165–170. Citado na página 56.

BIMSCHAS, D. et al. Semantic-service provisioning for the internet of things. In: EUROPEAN ASSOCIATION OF SOFTWARE SCIENCE AND TECHNOLOGY. *Workshops der wissenschaftlichen Konferenz Kommunikation in Verteilten Systemen*. [S.l.]: Electronic Communications of the EASST, 2011. v. 37, p. 12. Citado 4 vezes nas páginas 15, 31, 35, and 45.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, v. 5, n. 3, p. 1–22, 2009. Citado na página 35.

BOLLEGALA, D.; WEIR, D.; CARROLL, J. Cross-domain sentiment classification using a sentiment sensitive thesaurus. *Knowledge and Data Engineering, IEEE Transactions on*, v. 25, n. 8, p. 1719–1731, Aug 2013. ISSN 1041-4347. Citado 2 vezes nas páginas 28 and 40.

BOLLEN, J.; MAO, H.; PEPE, A. *Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena*. 2011. Disponível em: <http://aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2826/3237>. Citado na página 28.

BOULOS, M. N. K.; AL-SHORBAJI, N. M. On the internet of things, smart cities and the who healthy cities. *International journal of health geographics*, v. 13, n. 1, p. 10–072X–13–10, Mar 27 2014. Citado na página 24.

BOYLE, D.; YATES, D.; YEATMAN, E. Urban sensor data streams: London 2013. *Internet Computing, IEEE*, v. 17, n. 6, p. 12–20, Nov 2013. Citado 3 vezes nas páginas 29, 39, and 45.

BREITMAN, K.; CASANOVA, M. A.; TRUSZKOWSKI, W. *Semantic Web: Concepts, Technologies and Applications (NASA Monographs in Systems and Software Engineering)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 184628581X. Citado na página 35.

BUCCAFURRI, F. et al. A system for extracting structural information from social network accounts. *Software: Practice and Experience*, John Wiley and Sons, Ltd, v. 45, n. 9, p. 1251–1275, 2015. ISSN 1097-024X. Disponível em: <http://dx.doi.org/10.1002/spe.2280>. Citado na página 26.

BUENO JUAN ANTONIO ROJO, P. R. C. An experiment on semantic emotional evaluation of chats. In: *The Fifth International Conference on Advances in Semantic Processing*. [S.l.: s.n.], 2011. (SEMAPRO 2011), p. 116 to 121. ISBN 2308-4510. Citado na página 85.

CHEN, X.; VORVOREANU, M.; MADHAVAN, K. Mining social media data for understanding students x2019; learning experiences. *IEEE Transactions on Learning Technologies*, v. 7, n. 3, p. 246–259, July 2014. ISSN 1939-1382. Citado na página 26.

CHEN, X.-W.; LIN, X. Big data deep learning: Challenges and perspectives. *Access, IEEE*, v. 2, p. 514–525, 2014. ISSN 2169-3536. Citado na página 16.

CHEONG, M.; LEE, V. A study on detecting patterns in twitter intra-topic user and message clustering. In: *Pattern Recognition (ICPR), 2010 20th International Conference on*. [S.l.: s.n.], 2010. p. 3125–3128. ISSN 1051-4651. Citado 2 vezes nas páginas 26 and 27.

COLBAUGH, R.; GLASS, K. Analyzing social media content for security informatics. In: *Intelligence and Security Informatics Conference (EISIC), 2013 European*. [S.l.: s.n.], 2013. p. 45–51. Citado 2 vezes nas páginas 28 and 41.

COMAS, P. R.; TURMO, J.; MàRQUEZ, L. Sibyl, a factoid question-answering system for spoken documents. *ACM Trans. Inf. Syst.*, ACM, New York, NY, USA, v. 30, n. 3, p. 19:1–19:40, set. 2012. ISSN 1046-8188. Disponível em: <http://doi.acm.org.ez67.periodicos.capes.gov.br/10.1145/2328967.2328972>. Citado na página 56.

COMPTON, M. et al. A survey of the semantic specification of sensors. In: CEUR WORKSHOP PROCEEDINGS. *Proceedings of the 2nd International Workshop on Semantic Sensor Networks (SSN09) at ISWC 2009*. 2009. v. 522, p. 17–32. Disponível em: <http://ceur-ws.org/Vol-522/p7.pdf>. Citado na página 30.

CORCHO, O. *Semantics, Sensor Networds and Linked Stream/Sensor Data*. Cercedilla, 2011. Disponível em: <http://www.oeg-upm.net>. Citado 2 vezes nas páginas 45 and 52.

COSTA, R. A. G. da; CUGNASCA, C. E. Use of data warehouse to manage data from wireless sensors networks that monitor pollinators. In: *2010 Eleventh International Conference on Mobile Data Management*. [S.l.: s.n.], 2010. p. 402–406. ISSN 1551-6245. Citado na página 52.

CRUZ-BENITO, J.; GARCIA-PENALVO, F.; THERON, R. Defining generic data collectors for learning analytics: Facing up the heterogeneous data from heterogeneous environments. In: *Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on*. [S.l.: s.n.], 2014. p. 365–366. Citado na página 33.

DEBNATH, L. *Wavelet transforms and time-frequency signal analysis*. Boston: Birkhauser, 2001. ISBN 9780817641047, 0817641041. Citado na página 30.

DEVLIN, B.; ROGERS, S.; MYERS, J. Big data comes of age. In: . [s.n.], 2012. Disponível em: <http://www.ibm.com/big-data/us/en/downloads/>. Citado 4 vezes nas páginas 19, 40, 52, and 88.

DOHERTY, S. T.; LEMIEUX, C. J.; CANALLY, C. Tracking human activity and well-being in natural environments using wearable sensors and experience sampling. *Social Science and Medicine*, v. 106, p. 83 – 92, 2014. ISSN 0277-9536. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0277953614000756>. Citado na página 89.

ENDRES-NIGGEMEYER, B. The mashup ecosystem. In: ENDRES-NIGGEMEYER, B. (Ed.). *Semantic Mashups*. Springer Berlin Heidelberg, 2013. p. 1–50. ISBN 978-3-642-36402-0. Disponível em: <http://dx.doi.org/10.1007/978-3-642-36403-7_1>. Citado na página 32.

EVANS, D. The internet of things how the next evolution of the internet is changing everything. In: *Cisco Internet Business Solutions Group IBSG*. [s.n.], 2011. Disponível em:

<http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf>. Citado 5 vezes nas páginas 14, 15, 19, 20, and 88.

FACCHINETTI, A.; SPARACINO, G.; COBELLI, C. An online self-tunable method to denoise cgm sensor data. *IEEE Transactions on Biomedical Engineering*, v. 57, n. 3, p. 634–641, March 2010. ISSN 0018-9294. Citado na página 29.

FARIA, M. L. L. de. *Regras práticas para apresentação de páginas Web em dispositivos fixos e móveis.* 110 p. Dissertação (Mestrado), 2007. Citado na página 52.

FAUQUEX, M. et al. Creating people-aware iot applications by combining design thinking and user-centered design methods. In: *Internet of Things (WF-IoT), 2015 IEEE 2nd World Forum on*. [S.l.: s.n.], 2015. p. 57–62. Citado 2 vezes nas páginas 23 and 49.

FULTON, J. Emote api. In: . [s.n.], 2014. Disponível em: <http://www.emotivetext.com/>. Citado na página 85.

GANTI, R.; YE, F.; LEI, H. Mobile crowdsensing: current state and future challenges. *Communications Magazine, IEEE*, v. 49, n. 11, p. 32–39, November 2011. ISSN 0163-6804. Citado na página 24.

GHAG, K.; SHAH, K. Comparative analysis of the techniques for sentiment analysis. In: *Advances in Technology and Engineering 2013 International Conference on*. [S.l.: s.n.], 2013. p. 1–7. Citado 2 vezes nas páginas 43 and 89.

GIFFINGER, R. Smart cities Ranking of European medium-sized cities. *October*, v. 16, n. October, p. 13–18, 2007. ISSN 02642751. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S026427519800050X>. Citado na página 92.

GIOVANNELLA, C. et al. Scenarios for active learning in smart territories. *IxDeA*, v. 16, p. 7–16, 2013. Citado 2 vezes nas páginas 80 and 81.

GLOROT, X.; BORDES, A.; BENGIO, Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. In: *In Proceedings of the Twenty-eight International Conference on Machine Learning, ICML*. [S.l.: s.n.], 2011. Citado na página 15.

GOBBI, A.; SPINA, S. Smart cities and languages: The language network. *Interaction Design and Architecture(s)*, n. 16, p. 37–46, 2013. Citado 2 vezes nas páginas 23 and 24.

GUINARD, D. Mashing up your web enabled home enabling physical mashups in a web of things. In: *10th International Conference on Web Engineering*. [s.n.], 2010. Disponível em: <http://de.slideshare.net/misterdom/touch-the-web2010physicalhomemashups>. Citado na página 14.

HEBELER, J. et al. *Semantic Web Programming*. Chichester, West Sussex, Hoboken, NJ: John Wiley & Sons Inc., 2009. ISBN 978-0-470-41801-7. Citado na página 35.

HELLIWELL, R. L. J.; SACHS, J. *World Happiness Report 2016 Update*. 2016. 1 p. Disponível em: <http://worldhappiness.report/>. Citado na página 88.

HENNIG, P. et al. Accelerate the detection of trends by using sentiment analysis within the blogosphere. In: *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. [S.l.: s.n.], 2014. p. 503–508. Citado na página 26.

INTEL, B. R. L. *Intel Berkeley Research Lab - Lab Data*. 2016. 1 p. Disponível em: <http://db.csail.mit.edu/labdata/labdata.html>. Citado 2 vezes nas páginas 54 and 65.

IOVANOVICI, A.; PRODAN, L.; VLADUTIU, M. Collaborative environment for road traffic monitoring. *ITS Telecommunications (ITST), 2013 13th International Conference on*, p. 232–237, 2013. Citado na página 92.

JI, X. et al. Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining*, v. 5, n. 1, p. 1–25, 2015. ISSN 1869-5469. Disponível em: <http://dx.doi.org/10.1007/s13278-015-0253-5>. Citado na página 27.

JONES, N. Computer science: The learning machines. *Nature*, v. 505, n. 7482, p. 146–148, 2014. ISSN 0028-0836. Citado na página 16.

KABADAYI, S.; PRIDGEN, A.; JULIEN, C. Virtual sensors: abstracting data from physical sensors. In: *World of Wireless, Mobile and Multimedia Networks, 2006. WoWMoM 2006. International Symposium on a*. [S.l.: s.n.], 2006. p. 6 pp.–592. Citado na página 31.

KAISLER, S. H. et al. Advanced Analytics – Issues and Challenges in a Global Environment. In: *2014 47th Hawaii International Conference on System Sciences*. IEEE, 2014. p. 729–738. ISBN 978-1-4799-2504-9. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6758694>. Citado 2 vezes nas páginas 26 and 27.

KAMILARIS, A.; PITSILLIDES, A. Towards interoperable and sustainable smart homes. In: *IST-Africa Conference and Exhibition (IST-Africa), 2013*. [S.l.: s.n.], 2013. p. 1–11. Citado na página 15.

KAWTRAKUL, A. et al. Development of an information integration and knowledge fusion platform for spatial and time based advisory services: Precision farming as a case study. In: *Global Conference (SRII), 2014 Annual SRII*. [S.l.: s.n.], 2014. p. 241–248. Citado na página 33.

KENDA, K. et al. Mashups for the web of things. In: ENDRES-NIGGEMEYER, B. (Ed.). *Semantic Mashups*. Springer Berlin Heidelberg, 2013. p. 145–169. ISBN 978-3-642-36402-0. Disponível em: <http://dx.doi.org/10.1007/978-3-642-36403-7_5>. Citado 3 vezes nas páginas 29, 36, and 37.

KHAN, F. et al. Implementation of affective states and learning styles tactics in web-based learning management systems. In: *Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on*. [S.l.: s.n.], 2010. p. 734–735. Citado 2 vezes nas páginas 84 and 85.

KHAN, M. F.; KHAN, A.; KHAN, K. Efficient word sense disambiguation technique for sentence level sentiment classification of online reviews. In: . [S.l.: s.n.], 2013. v. 25. ISBN 10135316. Citado na página 41.

KNOERI, C.; STEINBERGER, J. K.; ROELICH, K. End-user centred infrastructure operation: towards integrated end-use service delivery. *Journal of Cleaner Production*, v. 132, p. 229 – 239, 2016. ISSN 0959-6526. Absolute Reductions in Material Throughput, Energy Use and Emissions. Disponível em: <http:

//www.sciencedirect.com/science/article/pii/S0959652615011701>. Citado 2 vezes nas páginas 23 and 49.

KOHONEN, T. The self-organizing map. *Neurocomputing*, v. 21, n. 1–3, p. 1 – 6, 1998. ISSN 0925-2312. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231298000307>. Citado na página 27.

KOHONEN, T. Essentials of the self-organizing map. *Neural Networks*, v. 37, p. 52 – 65, 2013. ISSN 0893-6080. Twenty-fifth Anniversay Commemorative Issue. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0893608012002596>. Citado na página 27.

KONYS, A. Knowledge-Based Approach to Question Answering System Selection. In: Nunez, M and Nguyen, NT and Camacho, D and Trawinski, B (Ed.). *COMPUTATIONAL COLLECTIVE INTELLIGENCE (ICCCI 2015), PT I*. [S.l.], 2015. (Lecture Notes in Artificial Intelligence, 9329), p. 361–370. ISBN 978-3-319-24069-5; 978-3-319-24068-8. ISSN 0302-9743. 7th International Conference on Computational Collective Intelligence (ICCCI), Madrid, SPAIN, SEP 21-23, 2015. Citado na página 76.

KUZNETSOV, V. A.; MOCHALOV, V. A.; MOCHALOVA, A. V. Ontological-semantic text analysis and the question answering system using data from ontology. In: *2016 18th International Conference on Advanced Communication Technology (ICACT)*. [S.l.: s.n.], 2016. p. 1–1. Citado na página 56.

LAENDER, A.; SILVA, A. da. Cooperative research on web data management at ufmg and ufam - a brief report. In: *Latin American Web Conference, 2008. LA-WEB '08.* [S.l.: s.n.], 2008. p. 144–150. Citado na página 33.

LAZARIDIS, I.; MEHROTRA, S. Capturing sensor-generated time series with quality guarantees. In: *Data Engineering, 2003. Proceedings. 19th International Conference on.* [S.l.: s.n.], 2003. p. 429–440. Citado na página 30.

LEVINE, D. S. Neural network modeling of emotion. *Physics of Life Reviews*, v. 4, n. 1, p. 37 – 63, 2007. ISSN 1571-0645. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1571064506000327>. Citado na página 28.

LI, T. et al. A survey on wavelet applications in data mining. *SIGKDD Explor.Newsl.*, v. 4, n. 2, p. 49–68, dec 2002. Disponível em: <http://doi.acm.org.oala-proxy.surrey.ac.uk/10.1145/772862.772870>. Citado na página 52.

LIMA, A. B. de; AMAZONAS, J. R. de A. *Internet Teletraffic Modeling and Estimation*. Wharton, TX, USA: River Publishers, 2013. ISBN 8792982107, 9788792982100. Citado na página 65.

LIU, B.; HU, M.; CHENG, J. Opinion observer: Analyzing and comparing opinions on the web. In: *Proceedings of the 14th International Conference on World Wide Web*. New York, NY, USA: ACM, 2005. (WWW '05), p. 342–351. ISBN 1-59593-046-9. Disponível em: <http://doi.acm.org/10.1145/1060745.1060797>. Citado na página 41.

LIYANAGE, C.; MARASINGHE, A. Planning smart meal in a smart city for a smart living. In: *Biometrics and Kansei Engineering, 2013 International Conference on.* [S.l.: s.n.], 2013. p. 166–171. Citado 2 vezes nas páginas 15 and 23.

MADRIA, S.; KUMAR, V.; DALVI, R. Sensor cloud: A cloud of virtual sensors. *Software, IEEE*, v. 31, n. 2, p. 70–77, Mar 2014. ISSN 0740-7459. Citado na página 31.

MAKAZHANOV, A.; RAFIEI, D. Predicting political preference of twitter users. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. New York, NY, USA: ACM, 2013. (ASONAM '13), p. 298–305. ISBN 978-1-4503-2240-9. Disponível em: <http://doi.acm.org/10.1145/2492517.2492527>. Citado na página 28.

MAMMERI, A. et al. North-american speed limit sign detection and recognition for smart cars. In: *Local Computer Networks Workshops (LCN Workshops), 2013 IEEE 38th Conference on*. [S.l.: s.n.], 2013. p. 154–161. Citado na página 15.

MAO, X.; LI, Z. Implementing emotion-based user-aware e-learning. In: *CHI '09 Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2009. (CHI EA '09), p. 3787–3792. ISBN 978-1-60558-247-4. Disponível em: <http://doi.acm.org/10.1145/1520340.1520572>. Citado na página 85.

MARSAL-LLACUNA, M.-L. City indicators on social sustainability as standardization technologies for smarter (citizen-centered) governance of cities. *Social Indicators Research*, v. 128, n. 3, p. 1193–1216, 2016. ISSN 1573-0921. Disponível em: <http://dx.doi.org/10.1007/s11205-015-1075-6>. Citado na página 23.

MARSAL-LLACUNA, M.-L.; FABREGAT-GESA, R. Modeling citizens' urban time-use using adaptive hypermedia surveys to obtain an urban planning, citizen-centric, methodological reinvention. *Time and Society*, v. 25, n. 2, p. 272–294, 2016. Disponível em: <http://tas.sagepub.com/content/25/2/272.abstract>. Citado na página 23.

MIKA, P.; ELFRING, T.; GROENEWEGEN, P. Application of semantic technology for social network analysis in the sciences. *Scientometrics*, v. 68, n. 1, p. 3–27, 2006. ISSN 1588-2861. Disponível em: <http://dx.doi.org/10.1007/s11192-006-0081-5>. Citado na página 50.

MIN, M. Data-driven database middleware for ubiquitous sensor networks. In: *Information Science and Applications (ICISA), 2013 International Conference on*. [S.l.: s.n.], 2013. p. 1–2. Citado na página 29.

MITCHELL, L. et al. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, Public Library of Science, v. 8, n. 5, p. 1–15, 05 2013. Disponível em: <http://dx.doi.org/10.1371%2Fjournal.pone.0064417>. Citado na página 51.

MOCANU, D. et al. The twitter of babel: Mapping world languages through microblogging platforms. *PLoS ONE*, Public Library of Science, v. 8, n. 4, p. 1–9, 04 2013. Disponível em: <http://dx.doi.org/10.13712Fjournal.pone.0061981>. Citado na página 27.

MOMEN, A.; AHMADI-NOUBARI, H.; MIRZAEE, A. A new denoising method in spread spectrum systems based on wavelet transform. *IEEE Canadian Conference on Electrical and Computer Engineering*, v. 1, p. 2127–2130, 2006. Citado na página 63.

MONTE, G. et al. Standard of things, first step: Understanding and normalizing sensor signals. In: *Industrial Electronics Society, IECON 2013 - 39th Annual Conference of the IEEE*. [S.l.: s.n.], 2013. p. 118–123. ISSN 1553-572X. Citado na página 29.

MOUSANNIF, H.; KHALIL, I. The human face of mobile. In: LINAWATI et al. (Ed.). *Information and Communication Technology*. Springer Berlin Heidelberg, 2014, (Lecture Notes in Computer Science, v. 8407). p. 1–20. ISBN 978-3-642-55031-7. Disponível em: <http://dx.doi.org/10.1007/978-3-642-55032-4_1>. Citado 3 vezes nas páginas 25, 41, and 42.

MURTAGH, F.; HERNANDEZ-PAJARES, M. The kohonen self-organizing map method: An assessment. *Journal of Classification*, v. 12, n. 2, p. 165–190, 1995. ISSN 1432-1343. Disponível em: <http://dx.doi.org/10.1007/BF03040854>. Citado na página 52.

NATI, M. et al. Smartcampus: A user-centric testbed for internet of things experimentation. In: *Wireless Personal Multimedia Communications (WPMC), 2013 16th International Symposium on*. [S.l.: s.n.], 2013. p. 1–6. ISSN 1347-6890. Citado na página 54.

NOGUEIRA, L. A.; AMAZONAS, J. R. de A. Improvement of the signal to noise ratio by means of wavelet modulation. In: *Revista de Investigaciones - Universidad del Quindío*. [S.l.: s.n.], 2012. p. 1–8. Citado na página 72.

ORTIGOSA, A.; MARTIN, J. M.; CARRO, R. M. Sentiment analysis in facebook and its application to e-learning. *Computers in Human Behavior*, v. 31, p. 527, 2014. Citado 4 vezes nas páginas 28, 85, 86, and 89.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, Now Publishers Inc., Hanover, MA, USA, v. 2, n. 1-2, p. 1–135, jan. 2008. ISSN 1554-0669. Disponível em: <http://dx.doi.org/10.1561/1500000011>. Citado 2 vezes nas páginas 40 and 41.

PERCIVAL, D. B.; WALDEN, A. T. *Wavelet Methods for Time Series Analysis*. [S.l.]: Cambridge University Press, 2000. Citado 2 vezes nas páginas 30 and 64.

PERERA, C. et al. Context aware computing for the internet of things: A survey. In: . [S.l.: s.n.], 2014. v. 16, n. 1, p. 414–454. ISBN 1553-877X. Citado 2 vezes nas páginas 44 and 45.

PEREZ-SANAGUSTIN, M. et al. Augmenting reality and formality of informal and non-formal settings to enhance blended learning. *Learning Technologies, IEEE Transactions on*, v. 7, n. 2, p. 118–131, April 2014. ISSN 1939-1382. Citado na página 80.

PHUOC, D. L.; PARREIRA, J. X.; HAUSWIRTH, M. Challenges in linked stream data processing: A position paper. In: *SSN*. [S.l.: s.n.], 2010. Citado na página 30.

PIETROPAOLI, B.; DOMINICI, M.; WEIS, F. Virtual sensors and data fusion in a multi-level context computing architecture. In: *Information Fusion (FUSION), 2013 16th International Conference on*. Istanbul: IEEE, 2013. p. 101–108. ISBN 9786058631113. Citado 2 vezes nas páginas 54 and 55.

PISKORSKI, J.; TANEV, H.; BALAHUR, A. Exploiting twitter for border security-related intelligence gathering. In: *Intelligence and Security Informatics Conference (EISIC), 2013 European*. [S.l.: s.n.], 2013. p. 239–246. Citado na página 51.

POIRIER, D. et al. Automating opinion analysis in film reviews: The case of statistic versus linguistic approach. In: . Dordrecht: Springer Netherlands, 2011. v. 45, p. 125–140. ISBN 9400717563, 9789400717565. Citado na página 40.

POKRIC, B.; KRCO, S.; POKRIC, M. Augmented reality based smart city services using secure iot infrastructure. In: *Proceedings of the 2014 28th International Conference on Advanced Information Networking and Applications Workshops*. Washington, DC, USA: IEEE Computer Society, 2014. (WAINA '14), p. 803–808. ISBN 978-1-4799-2653-4. Disponível em: <http://dx.doi.org/10.1109/WAINA.2014.127>. Citado na página 82.

POSLAD, S. et al. Using a Smart City IoT to Incentivise and Target Shifts in Mobility Behaviour–Is It a Piece of Pie? *Sensors (Basel, Switzerland)*, v. 15, n. 6, p. 13069–96, 2015. ISSN 1424-8220. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-84930960160{&}partnerID=tZOtx>. Citado na página 92.

PRIETO, B. et al. Analyzing a web-based social network using kohonen's som. In: _____. *Computational and Ambient Intelligence: 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 20-22, 2007. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 911–918. ISBN 978-3-540-73007-1. Disponível em: <http://dx.doi.org/10.1007/978-3-540-73007-1_110>. Citado na página 27.

RANTANEN, P.; SILLBERG, P. Event calendar for internet data sources. In: *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*. [S.l.: s.n.], 2014. p. 1035–1040. Citado 5 vezes nas páginas 19, 33, 37, 38, and 46.

RASTOGI, S.; SINGHAL, R.; KUMAR, R. Article: A sentiment analysis based approach to facebook user recommendation. *International Journal of Computer Applications*, v. 90, n. 16, p. 21–25, March 2014. Published by Foundation of Computer Science, New York, USA. Citado 3 vezes nas páginas 40, 41, and 43.

RICHTER, M.; FLCKIGER, M. *User-Centred Engineering: Creating Products for Humans*. [S.l.]: Springer Publishing Company, Incorporated, 2014. ISBN 3662439883, 9783662439883. Citado na página 23.

RIEDEL, M.; MEMON, A.; MEMON, M. High productivity data processing analytics methods with applications. In: *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*. [S.l.: s.n.], 2014. p. 289–294. Citado na página 33.

RODRIGUEZ, P.; ORTIGOSA, A.; CARRO, R. Extracting emotions from texts in e-learning environments. In: *Complex, Intelligent and Software Intensive Systems (CISIS), 2012 Sixth International Conference on*. [S.l.: s.n.], 2012. p. 887–892. Citado 2 vezes nas páginas 84 and 85.

ROSCIA, M.; LONGO, M.; LAZAROIU, G. Smart city by multi-agent systems. In: *Renewable Energy Research and Applications (ICRERA), 2013 International Conference on*. [S.l.: s.n.], 2013. p. 371–376. Citado na página 15.

ROSHANAEI, M.; MISHRA, S. Studying the attributes of users in twitter considering their emotional states. *Social Network Analysis and Mining*, v. 5, n. 1, p. 1–13, 2015. ISSN 1869-5469. Disponível em: <http://dx.doi.org/10.1007/s13278-015-0278-9>. Citado na página 26.

ROUTRAY, P.; SWAIN, C. K.; MISHRA, S. P. A survey on sentiment analysis. In: . [S.l.: s.n.], 2013. v. 76, n. 10. ISBN 0975-8887. Citado 4 vezes nas páginas 28, 40, 41, and 44.

RYOKAI, K. et al. Communicating and interpreting wearable sensor data with health coaches. In: *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2015 9th International Conference on*. [S.l.: s.n.], 2015. p. 221–224. Citado na página 88.

SABOSIK, P. Personalizing the internet of things. *Information Today*, v. 30, n. 11, p. 33, 2013. ISSN 87556286. Disponível em: <http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=bth&AN=92688532&site=ehost-live&custid=s4121186>. Citado 2 vezes nas páginas 17 and 33.

SAITO, S.; TOMIOKA, R.; YAMANISHI, K. Early detection of persistent topics in social networks. In: *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. [S.l.: s.n.], 2014. p. 417–424. Citado na página 26.

SALOUN, P.; HRUZIK, M.; ZELINKA, I. Sentiment analysis - e-bussines and e-learning common issue. In: *Emerging eLearning Technologies and Applications (ICETA), 2013 IEEE 11th International Conference on*. [S.l.: s.n.], 2013. p. 339–343. Citado na página 28.

SANTINI, S.; RöMER, K. An adaptive strategy for quality-based data reduction in wireless sensor networks. In: *in: Proceedings of the 3rd International Conference on Networked Sensing Systems (INSS'06*. [S.l.: s.n.], 2006. p. 29–36. Citado na página 30.

SEQUEDA, J. F.; CORCHO, O. *Linked Stream Data: A Position Paper*. 2009. 148–157 p. Citado na página 30.

SHETH, A.; ANANTHARAM, P.; HENSON, C. Physical-cyber-social computing: An early 21st century approach. *IEEE Intelligent Systems*, IEEE Computer Society, Los Alamitos, CA, USA, v. 28, n. 1, p. 78–82, 2013. ISSN 1541-1672. Citado na página 19.

SHETH, A.; HENSON, C.; SAHOO, S. S. Semantic sensor web. *IEEE Internet Computing*, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 12, n. 4, p. 78–83, jul. 2008. ISSN 1089-7801. Disponível em: <http://dx.doi.org/10.1109/MIC.2008.87>. Citado na página 30.

SHETH, A. P.; THIRUNARAYAN, K. Semantics empowered web 3.0: Managing enterprise, social, sensor, and cloud-based data and services for advanced applications. In: *Semantics Empowered Web 3.0*. [S.l.: s.n.], 2012. Citado 2 vezes nas páginas 34 and 47.

SHO, S.-H. et al. Ttcg: Three-tier context gathering technique for mobile devices. In: *Proceedings of the 5th International Conference on Pervasive Services*. New York, NY, USA: ACM, 2008. p. 157–162. ISBN 978-1-60558-135-4. Citado 2 vezes nas páginas 45 and 52.

SHULL, F. Getting an intuition for big data. *Software, IEEE*, v. 30, n. 4, p. 3–6, July 2013. Citado 5 vezes nas páginas 15, 19, 32, 45, and 88.

SPRAKE, J.; ROGERS, P. Crowds, citizens and sensors: process and practice for mobilising learning. In: . [S.l.: s.n.], 2014. v. 18, n. 3, p. 753–764. ISBN 1617-4909. Citado na página 24.

STRAPPARAVA, C.; MIHALCEA, R. Learning to identify emotions in text. In: *Proceedings of the 2008 ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2008. (SAC '08), p. 1556–1560. ISBN 978-1-59593-753-7. Disponível em: <http://doi.acm.org/10.1145/1363686.1364052>. Citado na página 85.

SUAREZ-TANGIL, G. et al. Evolution, detection and analysis of malware for smart devices. *Communications Surveys Tutorials, IEEE*, v. 16, n. 2, p. 961–987, Second 2014. ISSN 1553-877X. Citado na página 15.

TAN, S. et al. Interpreting the public sentiment variations on twitter. *Knowledge and Data Engineering, IEEE Transactions on*, v. 26, n. 5, p. 1158–1170, May 2014. ISSN 1041-4347. Citado 2 vezes nas páginas 28 and 51.

THOMAS, K. et al. Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse. In: *Presented as part of the 22nd USENIX Security Symposium (USENIX Security 13)*. Washington, D.C.: USENIX, 2013. p. 195–210. ISBN 978-1-931971-03-4. Disponível em: <https://www.usenix.org/conference/usenixsecurity13/technical-sessions/paper/thomas>. Citado na página 26.

TOMLJANOVIć, J.; PAVLIć, M.; KATIć, M. A. Intelligent question x2014; answering systems: Review of research. In: *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*. [S.l.: s.n.], 2014. p. 1228–1233. Citado na página 34.

TRSTENJAK, B.; MIKAC, S.; DONKO, D. Knn with tf-idf based framework for text categorization. *Procedia Engineering*, v. 69, p. 1356 – 1364, 2014. ISSN 1877-7058. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877705814003750>. Citado na página 51.

VARGAS-VERA, M.; LYTRAS, M. D. Aqua: A closed-domain question answering system. *Information Systems Management*, v. 27, n. 3, p. 217 – 225, 2010. ISSN 10580530. Disponível em: <http://search-ebscohost-com.ez67.periodicos.capes.gov.br/login.aspx?direct=true&db=iih&AN=52237303&lang=pt-br&site=ehost-live>. Citado na página 76.

VISSER, U. Question/answering systems. *KI - Kunstliche Intelligenz*, v. 26, n. 2, p. 191–195, 2012. ISSN 1610-1987. Disponível em: <http://dx.doi.org/10.1007/s13218-012-0172-9>. Citado na página 34.

WANG, Y. S.; HE, H.; CHEN, X. W. Comparison and application of signal denoising techniques based on time-frequency algorithms. In: *Intelligent Vehicles Symposium, 2009 IEEE*. [S.l.: s.n.], 2009. p. 129–133. ISSN 1931-0587. Citado na página 29.

WILSON, T.; WIEBE, J.; HOFFMANN, P. Recognizing contextual polarity in phrase-level sentiment analysis. In: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. (HLT '05), p. 347–354. Disponível em: <http://dx.doi.org/10.3115/1220575.1220619>. Citado na página 41.

YANAGIMOTO, H.; SHIMADA, M.; YOSHIMURA, A. Document similarity estimation for sentiment analysis using neural network. In: *Computer and Information Science (ICIS), 2013 IEEE/ACIS 12th International Conference on*. [S.l.: s.n.], 2013. p. 105–110. Citado na página 28.

YE, J.; DOBSON, S.; MCKEEVER, S. Review: Situation identification techniques in pervasive computing: A review. *Pervasive Mob. Comput.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 8, n. 1, p. 36–66, fev. 2012. Citado 2 vezes nas páginas 30 and 45.

YU, L.; LIU, Y. Using linked data in a heterogeneous sensor web: challenges, experiments and lessons learned. *International Journal of Digital Earth*, v. 0, n. 0, p. 1–21, 2011. Citado 2 vezes nas páginas 26 and 45.

ZERFOS, P. et al. Platform and applications for massive-scale streaming network analytics. *IBM Journal of Research and Development*, v. 57, n. 3/4, p. 11:1–11:13, May 2013. Citado na página 46.

ZHANG, D. G. et al. A new denoising algorithm based on spherical coordinates for iot. In: *Computer, Consumer and Control (IS3C), 2012 International Symposium on*. [S.l.: s.n.], 2012. p. 781–784. Citado na página 29.

ZHANG, Q.; TJHUNG, T. Outage Performance of Cellular Systems over Arbitrary Lognormal Shadowed Rician Channels. *IEE Electronics Letters*, v. 35, n. 15, p. 1227–1229, jul. 1999. Citado 2 vezes nas páginas 45 and 88.

ZHANG, W.; YOSHIDA, T.; TANG, X. A comparative study of tf*idf, {LSI} and multi-words for text classification. *Expert Systems with Applications*, v. 38, n. 3, p. 2758 – 2765, 2011. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417410008626>. Citado na página 51.

ZINCK, A.; NEWEN, A. Classifying emotion: a developmental account. *Synthese*, v. 161, n. 1, p. 1–25, 2008. Citado na página 85.

Appendix

APPENDIX A – Published Papers

1. Paper accepted by the IEEE Sensors Journal on 10 May 2016 and resubmitted on 08 July 2016:

    a) Title: Insights into IoT data and an innovative DWT-based technique to denoise sensor signals.

2. Paper accepted by the IEEE ISCE 2016 on 17 July 2016:

    a) Title: Using wearable devices to measure local happiness.

3. Poster accepted by the IEEE MASS 2016 on 15 July 2016:

    a) A platform for implementing the concept of citizen as a sensor.

4. Poster accepted for the MOBICOM / N2Women Workshop 06.09.2016

    a) Adding value to WSN reducing signal noises.

APPENDIX B – Poster Presentations

1. VII ABEP 2015 Conference – Imperial College London - 21/02/2015;
2. I APEB - NL - Annual Meeting and IV European Conference of Brazilian Students and Researchers, Utrecht University - Holland - 18/04/2015;
3. WPG-EC 2015 - IV Workshop de Pós-Graduação da Área de Concentração Engenharia de Computação do Programa de Pós-Graduação em Engenharia Elétrica da EPUSP - 15/10/2015.