UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

Tesi di Dottorato in Ingegneria Informatica e Automatica
XXV Ciclo

Semantic-based Knowledge Management and Document Processing in the e-Health domain

Author:
Dott.ssa Sara Romano

Supervisor:
Prof. Antonino Mazzeo
Coordinator:
Prof. Franco Garofalo

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

SecLab Group
Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione

April 2013

A computer would deserve to be called intelligent if it could deceive a human into believing that it was human.

Alan Turing

UNIVERSITY OF NAPLES *"FEDERICO II"*

Abstract

Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione

Doctor of Philosophy

Semantic-based Knowledge Management and Document Processing in the e-Health domain

by Sara ROMANO

Nowadays efficient knowledge management, organization and sharing has become a critical success factor. Due to the ease of data production within the Internet era, knowledge workers are increasingly overwhelmed by information from a bewildering array of information sources: emails, intranets, the web, social networks, etc. and yet still find it hard to access the specific information required for the task at hand. This implies that knowledge worker productivity is reduced and that organizations may be making decisions on the basis of incomplete knowledge. Social networking forms an important part of online activities of Web users and thus represent an information source able to provide useful information in real time. In recent years, Twitter, a microblogging service, has received much attention in research communities interest as a social medium for communicating with others and reporting news events. Thus in many contexts, as medical, juridical and humanistic ones, advanced knowledge management methodologies are needed to deal with the huge amount and heterogeneity of data. In recent years several applications raised in order to support operators, working within different sectors, across the life cycle of a digital document. What makes these instruments often unproductive is the fact that they were born to support the traditional, manual documentation process mainly based on the massive use of paper but, given the many technological and regulatory constraints, they can not completely replace the traditional paper process. For this purpose, the research activity I done in this work is aimed to investigate and propose knowledge management methodologies and techniques, mainly focused on the issues of information extraction, data mining and semantic document processing applied to heterogeneous and unstructured data.

Preface

Some of the research and results described in this Ph.D. thesis has undergone peer review and has been published in, or at the date of this printing is being considered for publication in, academic journals, books, and conferences. In the following I list all the papers developed during my research work as Ph.D. student.

1. Amato F., Gargiulo F., Mazzeo A., Romano S., Sansone C., *"Combining syntactic and semantic vector space models in the health domain by using a clustering ensemble"*. In Proceedings of the Sixth International Conference on Health Informatics (HEALTHINF 2013).

2. Amato F., Mazzeo A., Mazzocca N., Romano S.. *"A Semantic Document Composition SaaS: the CloSe System"*. Under minor review for International Journal of Computational Science and Engineering (IJCSE 2013).

3. Amato F., Mazzeo A., Mazzocca N., Romano S., *"CloSe: a Cloud SaaS for Semantic document composition"*. In the Sixth International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS 2012) Workshop on Semantic Web/Cloud Information and Services Discovery and Management (SWISM 2012).

4. Kanhabua N., Romano S., Stewart A., Nejdl W. *"Supporting Temporal Analytics for Health-Related Events in Microblogs"*. In Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM 2012).

5. Kanhabua N., Romano S., Stewart A., *"Identifying Relevant Temporal Expressions for Real-world Events"*. In the 35th International Conference of Special Interest Group of Information Retrieval (SIGIR 2012) Workshop on Time-aware Information Access (TAIA 2012).

6. Amato F., Casola V., Mazzocca N., Romano S.. *"A semantic approach for fine-grain access control of E-Health documents"*. Logic Journal of the IGPL first published online on August 3, (LJIGPL 2012).

7. Amato F., Casola V., Mazzeo A., Romano S.. *"A semantic based framework to identify and protect E-Health critical resources"*. In Journal on Information Assurance and Security, Volume 7, Issue 4, (JIAS 2012).

8. Amato F., Casola V., Mazzeo A., Mazzocca N., Romano S.. *"Approccio semantico per un trattamento documentale massivo in domini specialistici"*. In Proceedings of Conference on Smart Tech & Smart Innovation - la strada per costuire futuro (AICA 2011).

9. Amato F., Casola V., Mazzocca N., Romano S.. *"A semantic-based document processing framework: a security perspective"*. In the 5th International conference on Complex, Intelligent and Software Intensive Systems (CISIS 2011) Workshop on Semantic Web/Grid Information and Services Discovery and Management (SWISM 2012).

10. Amato F., Fasolino A.R., Mazzeo A., Moscato V., Picariello A., Romano S., Tramontano P.. *"Ensuring semantic interoperability for e-health applications"*. In the 5th International conference on Complex, Intelligent and Software Intensive Systems (CISIS 2011) Workshop on Semantic Web/Grid Information and Services Discovery and Management (SWISM 2012).

11. Amato F., Casola V., Mazzeo A., Romano S.. *"An innovative framework for securing unstructured documents"*. In proceedings of the 4th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2011).

12. Amato F, Mazzeo A, Romano S., Scippacercola S.. *"Evaluating peculiar Lexicon for Medical Record Sections Identification"*. In Proceedings of Innovazione e Societá (IES 2011).

13. Amato F., Mazzeo A., Romano S., Scippacercola S.. *"An iterative approach for lexicon characterization in juridical context"*. In proceedings of 7th Conference of the Italian Chapter of Association for Information Systems (itAIS 2010).

14. Amato F., Casola V., Mazzeo A., Romano S.. *"A semantic based methodology to classify and protect sensitive data in medical records"*. In proceedings of Sixth International Conference on Information Assurance and Security (IAS 2010).

15. Romano S., Cutugno F.. *"New Features in Spoken Language Search Hawk (SpLaSH): Query Language and Query Sequence"*. In proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010).

Contents

List of Figures

List of Tables

Abbreviations

AI	**A**rtificial **I**ntelligence
CS	**C**ognitive **S**ciences
DC	**D**ublin **C**ore
DE	**D**ocument **E**ngineering
DM	**D**ata **M**ining
DMS	**D**ocument **M**anagement **S**ystems
ED	**E**vent **D**etection
EHR	**E**lectronic **H**ealth **R**ecord
IE	**I**nformation **E**xtraction
IS	**I**nformation **S**tructuring
KM	**K**nowledge **M**anagement
NLP	**N**atural **L**anguage **P**rocessing
OCR	**O**ptical **C**haracter **R**ecognition
PPV	**P**redictive **P**ositive **V**alue
SW	**S**emantic **W**eb

Dedicated to my Family.

Chapter 1

Introduction

The high degree of specialization in our society, established by several technological developments ranging form written language, the printing press and finally to the Internet, has lowered the costs and accelerated the process of information distribution many orders of magnitude [148]. In the Internet era, large-scale computer networks and the pervasive World Wide Web infrastructure have largely solved the problem of providing ubiquitous access to any kind of information, allowing any party of the Internet global community to share information with any other party. Nowadays efficient knowledge management, organization and sharing has become a critical success factor. The knowledge is contained into packing units commonly named documents. Generally speaking the term *document* could be defined as "a writing containing information" [1] but the notion about the specific definition of document have been widely and continuously changing over time following the evolution of human and technological society. Depending on the viewpoints, the age and the working environment, the document definition ranged among storage medium to model for information interchange [33, 34, 119, 148]. The digital era facilitated the information production and changed completely the storage mean of documents from paper to electronic leading to a new viewpoint of its definition with a different concept of what constitutes a document. In the field of Computer Science a document is seen as a "file" containing data that could be used by applications. Based on these consideration, it is possible to conclude that whatever contains a written information could be considered as a document independently of the storage mean, the knowledge encoding and the structure. Thus documents are any kind of information content such as e-mails, web pages or microblogs and social media contents.

Due to the ease of data production within the Internet era, knowledge workers are increasingly overwhelmed by information from a bewildering array of information sources: emails, intranets, the Web, microblogs, etc. and yet still find it hard to access the specific information required for the task at hand. This implies that knowledge worker

productivity is reduced and that organizations may be making decisions on the basis of incomplete knowledge. Furthermore, an inability to access key information can lead to compliance failure [47, 48]. The technical and scientific issues related to this context have been designated as the "Big Data" challenges and have been identified as highly strategic by major research agencies. Social networking forms an important part of online activities of Web users and thus represent an information source able to provide useful information in real time. In recent years, Twitter, a microblogging service, has received much attention in research communities interest as a social medium for communicating with others and reporting news events. Given its real-time nature and volume (more than 140 millions of messages produced every day), Twitter data is also becoming a valuable source for applications dealing with trend detection, natural disaster detection and for public health in order to monitoring and detecting events [37, 40, 41, 126, 137, 144]. Thus in many contexts, as medical, juridical and humanistic ones, advanced knowledge management methodologies are needed to deal with the huge amount and heterogeneity of data.

These knowledge management issues have led to several research efforts addressing several aspects as *information extraction* (IE), *information categorization* and *access control policies*.

Information extraction is the task of automatically extracting structured information from unstructured and/or semi-structured documents exploiting different kinds of text analysis. Those are mostly related to techniques of Natural Language Processing (NLP) and to cross-disciplinary perspectives including Statistical and Computational Linguistics [32, 36, 90], whose objective is to study and analyze natural language and its functioning through computational tools and models. Informations are structured and represented by means of data representation models such as metadata. Metadata are commonly defined as *"data about data"* and represent a communications medium that meets the data structures heterogeneity and support interoperability. Several standards for data representation models have been developed ranging from general to domain specific formats [42, 57, 110, 127]. Moreover techniques of information extraction can be associated with text mining and semantic technologies activities in order to detect relevant concepts from textual data aiming at detecting events, indexing and retrieval of information as well as long term preservation issues [8].

Information categorization is the process of grouping data into categories or classes. Among others, it involves Classification and Clustering techniques that are respectively commonly known as supervised and unsupervised learning techniques for automatic document organization. Both analysis techniques represent the task of discover natural groupings (called clusters) of a set of object so that objects within a cluster should be as

similar as possible and objects belonging to one cluster should be as dissimilar as possible from objects belonging to other clusters. In the classification task part of data are labeled in order to apply the the algorithms and to detect groups of related data. The labeling process is done manually by human intervention. On the other hand, in clustering analysis there is no human expert who has assigned objects to classes but it is the distribution and makeup of the data that will determine cluster membership [27, 105]. In literature there are wide variety of different algorithms that solve the categorization task by means of classification or clustering techniques, each one differs from others in their notion of what constitutes a cluster and how efficiently find them. Moreover each algorithm rely on models of data organization and implementation may depend on the domain in which it should operate.
Regarding the *access control* issue, the access control models exploit the concept of data classification to protect critical resources. The Bell La Padula model [25] is a significant example of access control rules that are based on the security levels of the user-requestor and the resource-requested. Many solutions have been proposed in order to address document security and access [66, 68]. A quite new security research field in the literature is the adoption of semantic approaches in policy management [20].

The application of those issues include domains as the health one. The e-Health (Electronic Health) is going to change the way how patients and health care providers interact. The application of information technologies to the health care system has led to a growth of the health organization: multiple actors in the health care sector, with different interests, must be brought together to work towards a common goal in which the central position of patients within the care process is essential.
The challenge of e-Health is to contribute to good healthcare by providing value-added services to the health care actors (patients, doctors, etc...) and, at the same time, by enhancing the efficiency and reducing the costs of complex informative systems through the use of information and communication technologies. For health care providers, having rapid access to patient information is a critical aspect in delivering high quality care and managing costs. In Figure 1.1 is depicted an example of helthcare actors involved in the medical record management. The European Commission wants to boost the digital economy by enabling all Europeans to have access to on-line medical records anywhere in Europe by 2020. With the newly enacted Directive 2011/24/EU on patients' rights in cross-border health care due for implementation by 2013, it is inevitable that a centralized European health record system will become a reality even before 2020 [94]. With the European Commission pressures to make patient information available electronically not only to the providers, but to the patients, the problem of managing paper and electronic documents has never been greater.

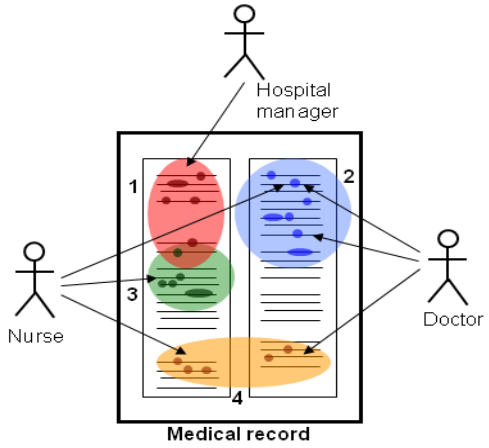

FIGURE 1.1: Healthcare actors involved in the medical record content management.

1.1 Thesis Contributions

In recent years several applications raised in order to support operators, working within different sectors, across the life cycle of a digital document, from its receipt until its processing and closure [6, 65, 113, 115]. What makes these instruments often unproductive is the fact that they were born to support the traditional, manual documentation process mainly based on the massive use of paper but, given the many technological and regulatory constraints, they can not completely replace the traditional paper process. Currently, there is a the strong coexistence of digital and paper-based information that makes the whole process expensive, unwieldy and partly fallacious. The reason is also related to the fact that numerous documents, although digital, are not structured at all causing several inconveniences for the automation of previously manual processes ranging from document management to determine fine-grained protection of the information. At the same time, during the processing and preparation of textual data, it comes from the need of retrieve and reuse information that could be complementary to the contents handled, to find a way to manage information coming from different sources and that could be presented in different forms (multimedia, web pages, social media, microblogs) and languages (multilingualism). Moreover, although documents are an established means of communication, their creation is costly, slow and not always needed. Often only small parts of a document are needed to answer a given information need [148].

For this purpose, the research activity I done in this work is aimed to investigate and propose knowledge management methodologies and techniques, mainly focused on the issues of information extraction, data mining and semantic document processing applied

to heterogeneous and unstructured data. To validate the results quality of the proposed methodologies, I answer to the following research question:

Question 1: Is it possible to automate processes for data classification of paper documents? And, how?

Question 2: Can be the access policies applied automatically for fine-grain protection of information?

Question 3: How external sources can be exploited for complementing traditional information?

To answer those questions, I proposed a reconfigurable framework for knowledge management and document processing that has been instantiated, applied and tested in the e-Health domain.
In the health domain, the information availability coming from different sources can improve the health services quality. For example, when a doctor has to deal with a diagnosis task, it could result very difficult to determine the patient's disease because of few and common nature of symptoms. Moreover a many diseases have several symptoms in common and knowing the place of origin of the disease can improve the diagnosis task and also the treatment assignment. For example, a disease that is common and frequent in a country can occur in a place in which is unusual and rare leading to a possible diagnosis mistake. In this scenario, the doctor's work and the patient's care could be improved through the use of methodologies and technologies that facilitate the actors of the medical domain in accessing health information from various sources such as, for example, medical records of hospital departments belonging to different countries and information about outbreaks currently occurring world wide. To this aim, the medical data must be properly organized and information from the Web must be conveniently filtered and monitored in order to automatically detect events related to infectious diseases currently occurring. In this work it has been defined an architecture for heterogeneous and multi-language data management, in order to support the actors in the medical domain to accessing and retrieving useful information. In particular, this architecture supports the user in the medical record composition allowing to:

- Structuring the medical records and identify the sections to associate automatically access policies [11–15];
- Organize previously scanned medical records according to the field of diagnosis (hospital departments such as surgery, cardiology, etc..) [16];
- Manage information from external sources (as Twitter) in order to identify outbreaks of epidemics [88, 89].

1.2 Thesis structure

The remainder of this thesis is:

- Chapter 2 describes an overview of the related research effort. In particular the research areas involved, in this work, are: Semantic Web, Document Engineering and Knowledge Management. Semantic Web is the research area providing standards, formalisms and languages that are exploited in this work for ontologies definition and reuse. Document engineering is the research area that deals with the designing of interfaces and models in order to enhance information exchange through different applications. It involves topics related to Information Structuring (IS), formalization and representation. In this work some of the approaches dealing with DE are exploited in order to enhance functionalities of the proposed architecture. Finally Knowledge Management is the research area that provides techniques and methodologies applied to data in order to acquiring, creating, organizing and sharing informations. It involves several disciplines ranging from Cognitive Sciences (CS) to Artificial Intelligence (AI) including activities of Information Extraction (IE), Event Detection (DM) and Data Mining (DM) that are those mostly involved in this work.

- Chapter 3 provides a formalization of the framework for document processing and knowledge management as well as the definition of the data and the of the semantic methodology adopted in this work. The framework takes in input textual information, transforms them and produce in output structured elements. The framework aims at analyze texts and, exploiting semantic based methodology, it automatically extract relevant information, concepts and complex relations, as events, organizing the not structured information in a structured fashion.

- Chapter 4 introduce the architecture, as multiple document processing framework instances, aimed at properly organize medical data and filter and monitor information from the Web in order to automatically detect events related to infectious diseases currently occurring. This architecture exploits semantic based methodology in order to process data and extract informations as concepts or complex relations in order to implement several functionalities. In this chapter I'll show how the framework for knowledge management and document processing has been instantiated and applied in the medical domain in order to provide the answers to the research questions of Section 1.1. In particular the functionalities provided by the architecture are based on; fine grain medical record protection (policy access); automatic classification of scanned medical records; event extraction from web sources in order to enhance medical domain actors' work.

- Chapter 5 describes the evaluation process of the architecture functionalities. Experimental settings and results obtained provide validation to the answers for the three research question.

- Chapter 6 conclude this work, giving some discussion and observations on the results obtained.

Chapter 2

Related Research Efforts

The work described in this thesis is concerned with three main research areas: Knowledge Management (KM), Document Engineering (DE) and Semantic Web (SW) (Figure 2.1).

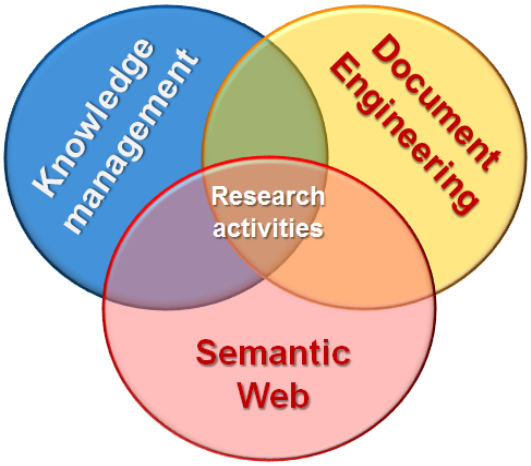

FIGURE 2.1: Research Areas

KM is the research area that provides techniques and methodologies applied to data in order to acquiring, creating, organizing and sharing informations. It involves several disciplines ranging from Cognitive Sciences (CS) to Artificial Intelligence (AI) including activities of Information Extraction (IE), Event Detection (DM) and Data Mining (DM) that are those mostly involved in this work. This is the research area whose state of the art is enhanced and for which related works provided could be compared to the approach described in this work.

DE is the research area that deals with the designing of interfaces and models in order to enhance information exchange through different applications. It involves topics related to Information Structuring (IS), formalization and representation. In this work some of

the approaches dealing with DE are exploited in order to enhance functionalities of the proposed architecture.

SW is the research area providing standards, formalisms and languages that are exploited in this work for ontologies definition and reuse. In this work, ontologies are used to enhance data structuring and interoperability.

Although those activities are general and could be applied to several domains, in this work the research effort is aimed to enhance methodologies and techniques of KM, DE and SW applied to the health domain. In this chapter is provided an overview of those three research areas. I will first outline the context domain followed by some characteristics of the SW that provided formalisms adopted in this work and finally by the DE and KM for which the state-of-the-art is enhanced in this thesis.

2.1 Contextualization: e-Health

The techniques and methodologies described above can be involved in different application domains as knowledge management and document processing have several basic issues in common within several areas as juridical, journalistic and administrative. In this work, those activities will be applied in the Health domain.

The e-Health (Electronic Health) is going to change the way how patients and health care providers interact. The application of information technologies to the health care system has led to a growth of the health organization: multiple actors in the health care sector, with different interests, must be brought together to work towards a common goal in which the central position of patients within the care process is essential.
The e-Health term encloses many meanings and services, ranging between medicine and information technologies. Just for example, emerging services are: the telemedicine (enhancing communication between doctors and patients by means of audiovisual media), the Consumer Health Informatics (optimizing the acquisition, storage, retrieval, and use of information in health), the m-Health (health care supported by mobile devices) and the Electronical Patient Records (improving patients health information sharing). The challenge of e-Health is to contribute to good healthcare by providing value-added services to the health care actors (patients, doctors, etc...) and, at the same time, by enhancing the efficiency and reducing the costs of complex informative systems through the use of information and communication technologies. For health care providers, having rapid access to patient information is a critical aspect in delivering high quality care and managing costs. The European Commission wants to boost the digital economy by enabling all Europeans to have access to on-line medical records anywhere in Europe

by 2020. With the newly enacted Directive 2011/24/EU on patients' rights in cross-border health care due for implementation by 2013, it is inevitable that a centralized European health record system will become a reality even before 2020 [94]. With the European Commission pressures to make patient information available electronically not only to the providers, but to the patients, the problem of managing paper and electronic documents has never been greater.

2.1.1 Electronic Health Records

The growth of information production and request leads to a revolutionary changing in the health care. For this reason, recently, the Electronic Health Records (EHR) have been introduced. The EHR is defined in [83] as *"digitally stored health care information about an individual's lifetime with the purpose of supporting continuity of care, education and research, and ensuring confidentiality at all times"*. The EHR contains several information including observations, clinical exams, treatments, therapies, administered drugs, allergies, patient vital statistics and legal information. Currently, this information are stored in healthcare available systems storage formats. Typically these formats are based on relational databases, structured documents and unstructured paper documents. As result, there is a lack of interoperability between systems belonging to the medical informatics. Interoperability aims to facilitate the interactions, allowing the information exchange and reuse, between non homogeneous information systems. Making EHR interoperable will lead to a more efficient health care services improving the retrieval and processing of patient's health information that can be stored in different sites. Transferring patient's information among several care sites will speed delivery and will reduce duplicate testing. Moreover, the possible introduction of automatic alerts will reduce human errors and will improve the benefit of patient's care. So, with the introduction of interoperability, the healthcare systems will be able to cooperate and sharing data and services without human extra intervention. The use of EHR also leads to some disadvantages as effort increase for the implementation, maintenance and use. These factors include the changing and the reconfigurations of EHR equipment. Paper health records are simpler to update and change. In fact, in this case, changing are simply treated by adding a new paper module into the medical record and the informations are stored in informal fashion. This is a useful characteristic in medical domain where the information structures are continuously evolving. So, modifying a paper health record is an easy task with respect to the electronic ones in which changes are expensive, slow to implement and can be unsatisfactory for the end users. Unfortunately, the data stored in paper module don't support interoperability. For example, these kinds of data are not

automatically processable leading to difficulties like collaboration supports with other medical structures or support for automatized decisions.

2.2 Semantic Web

In the early 2000s Tim Berners-Lee, the inventor of the World Wide Web and the director of the World Wide Web Consortium ("W3C"), laid down the foundations of a new form of Web content that is meaningful to computers and that is intended to unleash a revolution of new possibilities: the Semantic Web (SW) [28]. He defined the Semantic Web as *"an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"*.

The SW is an attempt to extend the potency of the Web with an analogous extension of peoples behavior enhancing the developing of software agents that include in part the human ability to process and determine the meaning of Web resources. This ability could be reached by enriching the Web resources with additional machine-readable descriptions of the contents (metadata) that complement the human-readable ones. The real success of the SW crucially depends on easy creation and management of semantic metadata by mass collaboration, i.e. by combining semantic content created by a large number of people [149]. It tries to get people to make their data available to others, and to add links to make them accessible by link following. So the vision of the SW is as an extension of Web principles from documents to data [29, 78]. Besides describing the available resources with metadata, one of the core challenges of the Semantic Web is concerned about making data and metadata to be efficiently shared, integrated and reused across application, enterprise, and community boundaries, as well as providing the agency to manage them. This creates the requirement that software agents must be able to process together data in heterogeneous formats, gathered using different principles for a variety of primary tasks. The Web's power will be that much greater if data can be defined and linked so that machines can go beyond display, and instead integrate and reason about data across applications (and across organizational or community boundaries) [29]. In the rest of this section the main components of the Semantic Web will be outlined.

2.2.1 Semantic Web Layer Cake

The original vision of the SW is encapsulated into a set of layered specifications and components known as Semantic Web Layer Cake [29]. The main concepts underlying the layer cake are a set standard technologies that are hierarchically organized to make

the Semantic Web possible. In the following I'll describe the main components belonging to the cake layer by bottom-up approach.

URI is an acronym for Uniform Resource Identifier and represent a compact string of characters used to identify a resource in the Web. An URL of a web site is a popular example and it could be considered as a subset of URI. Associating a URI with a resource enable the possibility to create links to it, refer to it or retrieve a representation of it. Relations, identified by URIs, link resources which are also identified by URIs.

XML (EXtensible Markup Language) [155] represent the syntactic unit of the languages for the SW that belongs to higher levels in the layer cake. Given its main characteristics of extensibility and self-describing it is widely accepted and used markup language.

The Resource Description Framework (RDF) [106] is the language used to describe the resources and their properties in order to allow software agents to "understand" and to process them. The elements that define this language are triples of *subject*, denoting the resource (URI), *predicate*, denoting properties and/or relationships of the resource, *object*, denoting the value of the property. Since RDF, provides no mechanisms for describing the properties (predicates), nor does it provide any mechanisms for describing the relationships between these properties and other resources, the RDF Schema was introduced. It provides mechanisms to define classes and properties that may be used to describe classes, properties and other resources. RDF and RDFS provide a standard domain-neutral model (mechanism) to describe individual resources. The model neither defines the semantics of any application domain, nor makes assumptions about a particular domain. Defining domain-specific features and their semantics requires additional facilities.

Ontologies are above RDF and RDFS. Ontologies are commonly defined as a *"specification of a conceptualization"* [76]. An ontology is a formal description of the concepts and relationships that are needed to understand a domain, and the vocabulary required to enter into a discourse about it, and how those concepts and vocabulary are interrelated, how classes and instances and their properties are defined, described and referred to [29]. The role of ontologies in the SW is mainly related to interoperability since they enable Web-based knowledge processing, sharing, and reuse among applications. The ontologies are encoded by means of OWL, Ontology Web Language, that provides a set of XML elements and attributes, with well-defined meanings, which are used to describe domain concepts and their relationships in an ontology.

The top layer in the SW architecture is composed by the Logic. Having described the Web content and published the ontological metadata, the next step is to discovery and use the semantics. The basic mechanism is to query for seek information that fulfills

explicit requirements. SPARQL [122] is a W3C recommended query language that is commonly accepted in the Semantic Web community. It enables the interrogation of amalgamated datasets to provide access to their combined information.

2.3 Document Engineering

An efficient knowledge organization and sharing has become one of the key to success in several professional and social activities. Technological developments like written language, the printing press and finally the Internet have lowered the costs and accelerated the process of information distribution many orders of magnitude. The knowledge is contained in *documents* that represent the packing format for spreading information [148]. The opinion about the definition of document have been widely and continuously changing over time following the evolution of human and technological society. Depending on the viewpoints, the age and the working environment, the document definition ranged among storage medium to model for information interchange [33, 34, 119, 148]. One of the pioneers in providing a definition of a document was Briet that in 1951 denoted it as *"any physical or symbolic sign, preserved or recorded, intended to represent, to reconstruct, or to demonstrate a physical or conceptual phenomenon"* [33]. The digital era changed completely the storage mean of documents from paper to electronic leading to a new viewpoint of document definition with a different concept of what constitutes a document. Buckland started the evolutionary change of digital document definition arguing that it should be considered in terms of function rather than physical function [34]. In [119] three different definition of the term document are provided. Document as a form, that is a container assembling data content and structure in order to make it readable both by its designer and its readers; document as a sign, meaning that the text should be processable by a knowledge system; document as a medium, where it is seen as a means of information distribution even in the future. In [148] the document is seen as a knowledge artifact consisting of several layers built on the top of information atoms (the words). Those layers determine the characteristics of a document ranging from its structure to the semantic of the content.

The main differences between paper and digital format of documents are the storage medium and the processes of creating, maintaining and preserving. In fact, with respect to paper, electronic documents take advantages of their compact storage, simple and fast updates and transmission, easy retrieval. Of course, in order to benefit of those advantages a set of programs designed to manage digital documents are required. Those programs are known as Document Management Systems (DMS). The DMS take advantages of metadata and informations or annotations connected to the documents in order

to provide the functionalities of storage security, indexing and retrieval. In the next sections the state-of-the-art of current metadata and some examples of available DMS are provided.

2.3.1 Metadata as annotation model of information

In the field of data modeling, several standards have been developed in order to *(i)* support interoperability and *(ii)* meet the data structures heterogeneity. The information contained in documents has been traditionally managed through the use of metadata. The most common definition of metadata is *"data about data"* and can be considered as a communications medium expressing semantics of informations in order to improve data retrieval [134]. Metadata can be defined and expressed in several languages and forms. It is possible to include Semantic Web Technologies in order to create metadata as ontologies that specify characteristics and content of documents. Dublin Core (DC) [42] is an ontology widely used to describe characteristics of digital resources exploiting a set of concepts as author, date of creation, format. Multimedia Content Description Interface (MPEG-7) [127] is a standard providing a set of description tools, for the symbolic description of documents that and multimedia (audio-video) contents.

2.3.1.1 Metadata and standards in the Health Domain

In the health domain several metadata and standards have been proposed as Healt Level 7 (HL7) Clinical Document Architecture (CDA) [57] and CEN EN 13606 EHRcom [110]. These standards aim to structure medical record contents for data exchange improvement. The IHE (Integratine the healthcare Enterprise)[1] initiative specify the Cross-Enterprise Document Sharing [143] standard to manage documents sharing between several healthcare organizations. The IHE Cross-Enterprise Document Sharing basic idea is to preserve the health document in a XML-based format in order to facilitate the sharing. A medical record may also contain images as for example from X-rays; DICOM (Digital Imaging and Communication in Medicine) [55] has become the de-facto standard for communication of medical images. This standard defines the data structures to facilitate the exchange of medical images and attached information. Of course there are also proposals to convert one standard to another. For example, the HL7 consortium proposes a mapping between DICOM SR "Basic Diagnostic Imaging Report" in HL7 CDA Release2 "Diagnostic Imaging Report" Mapping [56].

[1] Integrating the Healthcare Enterprise. http://www.ihe.net/

An important initiative called Good Electronic Health Record (GEHR) introduced the openEHR [114], a virtual community working on interoperability and computability in e-health. Its main focus is Electronic Health Records (EHRs) and systems. The OpenEHR Foundation that has published a set of specifications defining a health information reference model, a language for building "clinical models", or archetypes, which are separate from the software, and a query language. An archetype is a formal expression of a single concept such as, for example, "blood pressure", "laboratory results", "clinical exams" that are expressed as constraints on data whose instances conform to a reference model [23]. Components and systems conforming to openEHR are "open" in terms of data (they obey the published openEHR XML Schemas), models (they are driven by archetypes, written in the published ADL formalism) and APIs. They share the key openEHR innovation of adaptability, due to the archetypes being external to the software, and significant parts of the software being machine-derived from the archetypes. Systems based on archetypes specify standards for access to medical information exchange protocols and thus promoting information interoperability and accessibility. In order to meet future requirements, this standard has been designed so that it can be easy to expand it. In this way, the information contained in systems based on archetypes can be used across several institutions both at present and in the future.

Despite efforts to find a common standard for medical documents structuring and to facilitate the interoperability of information, many goals remain unfulfilled. The representation models of clinical information do not yet have a theoretical base strong enough to ensure information interoperability and computability. A model for the EHR should satisfy a large set of requirements including: computational efficiency, maintainability, scalability and extensibility requirements of the system for health information privacy and security. To meet these needs, in openEHR a new aspect was introduced: ontologies [114]. Ontologies are a formal way to describe aspects of a domain. These are used primarily for two reasons: a) people and machines can agree on the "facts" of the domain and b) inferences can be performed, usually based on the classification of "facts" in individual medical categories (e.g. the patient has a chronically high blood pressure means that the person is hypertensive) and alert classes (patient A has a high risk of stroke). As regards the first aspect (a) POMR Ontology (Problem-Oriented Record Ontology)[2] considers that a medical record is a repository of medical information and is the means of communication. POMR Ontology is an ontology that describes the medical records so that there is a unique vocabulary for electronic health records. As regards the second aspect (b) Beale and Heard [22] propose a model for clinical information based on health care ontological analysis seen like a problem-solving process. According to their

[2]http://esw.w3.org/HCLS/POMROntology

point of view, medical records contain a list of events, situations, etc. that are interpreted by professionals. The implication is that any model for the health information representation should be, in some way, the "cognitive" communication process of health professionals. To achieve this, the authors propose an ontology whose main purpose is to codify some types of information such as medical advice and observations from which the system is able to automatically identify actions that should be undertaken on the patient.

In order to address security and access control for EHR systems, several solutions have been proposed [66]. Although these solutions use role based access control for security management none of these took into account the structure and the semantics of EHRs. A first step in this direction was made in [86]. This approach focuses on identifying and organizing EHRs by means of semantic interpretation of internal data so that access control policies can be specified to authorize EHRs portions data sharing.

2.3.2 Document Management Systems

Nowadays organizations rely on computer databases to compile and preserve documents and private information, as for example human resources and finance within departments or agencies dealing with the environment or social care. The need to properly manage and preserve records for quality assurance implies that generic processes for electronic document and records management are required. These processes should essentially deal with capturing, classifying, indexing, retrieving and using information in collaborative framework together with their archival and disposal. The typical document life-cycle is depicted in Figure 2.2.

For this reason a number of private companies and software communities began to develop Document Management Systems (DMS) capable to deal and manage with all features involved in the electronic documents life-cycle. In this section some examples of DMS are outlined. This section is intended to provide a list of some existing commercial and open source systems available and to outline some functionalities feasible.

Alfresco [6] is an open Source Enterprise Content Management (CMS) including Web Content Management. Alfresco is the open platform for business critical document management and collaboration that automates document-intensive business processes and enables large-scale collaboration. OpenDocMan [113] is an open source document management system, written in PHP and it runs inside any popular web server. It was created to help companies with document management requirements. It is currently in active development, with new features being released regularly. OpenKM [115] is a open source electronic document management system with a web user interface that

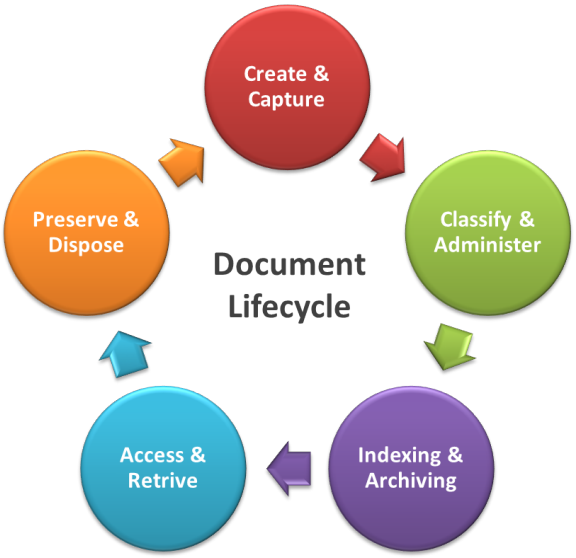

FIGURE 2.2: Document Life-cycle

allows the following operations to be carried out: sharing, setting security roles, auditing and finding enterprise documents and registers. ManagePoint [104] is a document management software that centralizes a company's documents, allowing instant access to information. WinDream [153] is a document management system that integrates into Windows and allows rich indexing and search options. eDocXL [62] is a scanning and document management software, which enables a user to scan, import and file any type of document, including paper, business cards, PDFs, photos, images, graphics and anything created by any Windows application. Empolis [65] offers an integrated suite of business applications which utilize semantic technologies for the analyzing, interpreting and processing of unstructured data.

Some examples of DMS developed for medical domain are: QuadraMed [123] an Electronic Document Management solution for your health care network that acts as a repository for downloaded and scanned documents. SoftTech Health [138] offers automated control of standard operating procedures and other files through automatic reviewing, archiving, conversion and delivery. KeyMark [91] provides records management systems and forms processing software for healthcare and other domains as insurance, entertainment and human resources.

2.3.3 Limitations of Document Management Systems

There are many applications that are springing up in recent years to support operators in different sectors across the life cycle of a digital document, from receipt processing, until its closure. What makes these instruments often unproductive is the fact that they

were born to support the documentation process traditional, manual, mainly based on the massive use of paper but, given the many technological and regulatory constraints, they can not completely replace the traditional paper process. The current state of the strong coexistence of digital information and paper-based information that makes it not only expensive and unwieldy the whole process, but make it partly fallacious. The reason for not efficiency is also related to the fact that numerous documents, although digital, are not structured at all and this causes several inconveniences for the activities of the automation of various processes of management and protection of the fine-grained information. At the same time, during the processing and preparation of textual data is becoming more and more the need to find and reuse information that can be complementary to these topics and that can come from different sources and present, therefore, in different forms (multimedia, unstructured text). In addition, the delivery of information, due to the globalization information itself, not more framed within predefined boundaries, leads to the necessity of dealing with data that may be in the form of encodings of different language (multilingualism).

2.4 Knowledge Management

In this section some related works dealing with KM will be provided. KM is a discipline widely adopted in several contexts due to the different meanings that the term *"knowledge"* takes when applied within a specific field.

The term *"knowledge"* founds place in many disciplines ranging from Cognitive Science to Artificial Intelligence, from Philosophy to Computer Science. Due to its universality, it has been defined differently depending on the field involved. Thus Knowledge is "Knowledge is the perception of the agreement or disagreement of two ideas" [103] and "Knowledge is a fluid mix of framed experience, contextual information, values and expert insight that provides a framework for evaluating and incorporating new experiences and information" [61] and another "Knowledge is richer, more structured and more contextual form of information" [93].
Part of the difficulty of defining knowledge arises from its relationship to two other concepts: *data* and *information*. Although these two terms have multiple definition, they are regarded as lower denomination of Knowledge. Thus the correlation between knowledge, information and data is encapsulated into layers, described as follows:

Data. Data represent the lower layer. These are facts and description of something specific belonging to the world, that are unstructured and not organized in any way. Thus data are not able to transmit informations about context and patterns of the facts and object described.

Information. On the top of data lies the information layer that capture and contextualize data.

Knowledge. Knowledge is the map of the world and it is a product of individual experience. Like a physical map, it helps us to know where things are containing also our beliefs and expectations. Knowledge can be classified as tacit or explicit. Tacit knowledge is within the individuals and cannot be reduced to digital form. It is generated by best practices and experience. However, it expresses in the social realm as the response ability of individuals (productivity, innovation and initiative), and teamwork (communication, coordination and collaboration). Explicit knowledge can be recorded digitally in documents, records, patents and other intellectual property artifacts. Explicit knowledge is representational and can live and be manipulated within the digital domain. Converting data-to-information and information-to-knowledge describes a value continuum of explicit knowledge [84].

In Computer Science, KM involves the integration of knowledge into computer systems in order to interpret and solve complex problems that normally require human competences and expertise [70]. Due to the ease of data production, within the Internet era knowledge workers are increasingly overwhelmed by information from a bewildering array of information sources: emails, intranets, the web, etc. and yet still find it hard to access the specific information required for the task at hand. This implies that knowledge worker productivity is reduced and that organizations may be making decisions on the basis of incomplete knowledge. Furthermore, an inability to access key information can lead to compliance failure [47, 48].
KM systems should provide instruments for building, maintaining and development of knowledge base which represent the central unit of any knowledge-based intelligent system. The first step involved in the knowledge base development consists in the activity, named knowledge acquisition, of automatically acquiring human knowledge and codifying it using appropriate representation formalisms and languages (as described in Section 2.2.1). Knowledge acquisition is strongly related to Information Extraction (IE) and Machine Learning techniques. Moreover semantic technology could be used to enhance intelligent information access facilities by annotating documents and informations with semantic meta-information. This allows more sophisticated analysis of information: for example, named entity recognition is a language processing technique which can identify particular locations, organizations or people mentioned in texts with ontological descriptions of those entities. Similarly, knowledge discovery techniques can be used to analyze content and classify it against an ontology, or indeed to derive new ontologies from content [47]. In this section I will provide related work dealing with

Knowledge Management aspects correlated to the work done in this thesis and related, in particular, to information extraction and information categorization applied to heterogeneous data from the Wed as well as to documents. As described in the next chapters, those activities were applied for event extraction, document classification and automatic access control policies association.

2.4.1 Information Extraction

Information extraction (IE) is the process of automatically scanning text for information relevant to some interest, including extracting entities, relations, and, most challenging, events (something happened in particular place at particular time) [79, 111]. It makes the information in the text more accessible for further processing. The increasing availability of on-line sources of information in the form of natural-language texts increased accessibility of textual information. The overwhelming quantity of available information has led to a strong interest in technology for processing text automatically in order to extract task-relevant information [17, 75, 111]. IE main task is to automatically extract structured information from unstructured and/or semi-structured documents exploiting different kinds of text analysis. Those are mostly related to techniques of Natural Language Processing (NLP) and to cross-disciplinary perspectives including Statistical and Computational Linguistics [32, 36, 90], whose objective is to study and analyze natural language and its functioning through computational tools and models. Moreover techniques of information extraction can be associated with text mining and semantic technologies activities in order to detect relevant concepts from textual data aiming at detecting events, indexing and retrieval of information as well as long term preservation issues [8]. Standard approaches used for implementing IE systems rely mostly on:

- Hand-written regular expressions. Hand-coded systems often rely on extensive lists of people, organizations, locations, and other entity types.
- Machine Learning (ML) based Systems. Hand annotated corpus is costly thus ML methods are used to automatically train an IE system to produce text annotation. Those systems are mostly based on supervised techniques to learn extraction patterns from plain or semi-structured texts. It is possible to distinguish two types of ML systems:
  - Classifier based. A part of manually annotated corpus is used to train the IE system in order to produce text annotation [95, 125, 136].
  - Active learning (or bootstrapping). In preparing for conventional supervised learning, one selects a corpus and annotates the entire corpus from beginning

to end. The idea of active learning involves having the system select examples for the user to annotate which are likely to be informative which are likely to improve the accuracy of the model [75]. Some examples of IE systems are [4, 117].

Various external knowledge sources have been successfully used to improve the IE process. They range from knowledge bases specific to a particular domain to general purpose resources applicable in many domains. One of the most widely used linguistic resources is the lexical network called wordnet [108]. It is a thesaurus which organizes the lexical units (literals) into synsets (groups of synonyms) and links them using various types of relations. Other general-purpose knowledge sources that are widely used include DOLCE[3] and SUMO[4].

There are several available software platforms, libraries and web services that can be directly applied for various IE tasks. Some examples are outlined in the following. The General Architecture for Text Engineering (GATE) is an open source platform for natural language processing implemented in Java. The Unstructured Information Management Architecture (UIMA) is a framework for integrating components processing any kind of unstructured information, such as text or multimedia. MinorThird, a set of Java classes for entity recognition based on ML methods. KNIME [30] is a user-friendly graphical workbench for the entire analysis process: data access, data transformation, initial investigation, visualisation and reporting. Weka [150] (http://weka.sourceforge.net) is another Java library implementing many general ML methods for classification, clustering as well as several NLP modules. Taltac [131] is a software for automatic analysis of italian text in the dual logic of text analysis and text mining. OpenCalais[112] is a publicly available web service aimed at extracting named entities, relations and events.

2.4.2 Information Categorization

The task of grouping informations and assigning a document into a class or category is known as Information Categorization. It involves Classification and Clustering techniques that are respectively commonly known as supervised and unsupervised learning techniques for automatic document organization. Both analysis techniques represent the task of discover natural groupings (called clusters) of a set of object so that objects within a cluster should be as similar as possible and objects belonging to one cluster should be as dissimilar as possible from objects belonging to other clusters.

[3] http://www.loa-cnr.it/DOLCE.html
[4] http://suo.ieee.org/SUO/Evaluations/

2.4.2.1 Supervised Learning: document classification

In the classification task part of data are labeled in order to apply the the algorithms and to detect groups of related data. The labeling process is done manually by human intervention [27, 105]. In literature there are wide variety of different algorithms that solve the categorization task by means of classification or clustering techniques, each one differs from others in their notion of what constitutes a cluster and how efficiently find them. Moreover each algorithm rely on models of data organization implementation may depend on the domain in which it should operate.

2.4.2.2 Unsupervised Learning: document clustering

Cluster analysis represents the task of discover natural groupings (called clusters) of a set of object so that objects within a cluster should be as similar as possible and objects belonging to one cluster should be as dissimilar as possible from objects belonging to other clusters. Clustering analysis is considered the most common form of unsupervised learning. It means that, differently from the classification task (that is a form of supervised learning), there is no human expert who has assigned objects to classes but it is the distribution and makeup of the data that will determine cluster membership [27] [105]. In fact, in the classification task a part of data should be labeled in order to apply the the algorithms and to detect groups of related data. Usually, this human intervention is not possible when the methods should be applied to a on-line systems dealing with different kind of data.
Cluster analysis should not be considered as a specific algorithm but it represents a task that should be solved. In literature there are wide variety of different algorithms that solve the clustering task, each one differs from others in their notion of what constitutes a cluster and how efficiently find them. It is clear that, since clustering analysis is used in several field, each algorithm implementation may depend on the domain in which it should operate. Moreover each clustering algorithm carry out its task according to the cluster model they implements. Typical cluster models are: connectivity models (hierarchical clustering) [100], centroid models (k-means) [52], distribution models (expectation-maximization) [51] and so on.

In general, cluster analysis is not an automatic task but an iterative process of knowledge discovery. The appropriate clustering algorithm (cluster model) and parameters settings, including the distance function, a density threshold, the number of expected clusters) depend on the particular problem to solve and thus on the considered domain, the data set and on the intended use of the results. For these reasons it can often be necessary to modify data preprocessing and parameters settings until the results obtained

have the desired properties. An important input parameter of each clustering algorithm is represented by the distance measure. There are several distance measure that can be used (as, for example, euclidean distance, Manhattan distance, power distance) each one strongly influence the outcome of clustering results.
Document clustering is closely related to the data clustering. It aims to discover natural groupings, and thus present an overview of the classes (topics) in a collection of documents. It is widely used in search engines for automatically grouping the retrieved document into a list of meaningful categories, as is achieved by Enterprise Search engines such as Vivisimo[5] or open source software such as Carrot2[6]. A good document clustering can be considered as the one that partition a documents collection into groups such that the elements within each group are both similar to each other and dissimilar to those in other groups. In order to determine what constitutes a good clustering there have been several evaluation metrics suggestions for a measure of similarity between two clustering. These measures can be divided into clustering internal evaluations, when a clustering result is evaluated based on the data that was clustered itself, and external evaluations, when clustering results are evaluated based on data that was not used for clustering, such as known class labels [72]. The internal evaluation metrics include DaviesBouldin index [46] and Dunn index [60] while the external evaluation metrics include Precision, Recall, and F-measure [105]. There have been several suggestions for a measure of similarity between two clusterings. Such a measure can be used to compare how well different data clustering algorithms perform on a set of data. One of the most important issues to be considered when dealing with a document clustering problem is to determine which features of a document are to be considered discriminatory. Many existing clustering approaches choose to represent each document as a vector where the attributes are terms (bag of words), therefore reducing a document to a representation suitable for traditional data clustering approaches [74, 96, 129]. Recently, several document clustering methods have been proposed. For example, different clustering algorithms where applied for the task of clustering multi-word terms in order to reflect a human-built ontology [130]; clustering algorithms were applied on a phrase graph model for determining document similarity [77]. Moreover the use of external resources for unsupervised learning purpose were proposed [38, 81, 135]. For example, Wikipidia [152] was used in order to represent a concept model for address the document clustering problem [81]; external ontologies were used for clustering and their request in order to improve semantic interoperability between companies and customers [135]; a gene ontology was used to infer similarity metric that was applied to detect sets of related genes with biological classifications [38].

[5]http://vivisimo.com/
[6]http://project.carrot2.org/

2.4.3 Event Detection

Event detection is an interesting task for many applications, for instance: surveillance, scientific discovery, and Topic Detection and Tracking. Numerous works have focused on detecting events from unstructured text and determining what features constitutes an event, e.g., key terms or named entities. Web resources are now considered as a valuable source for event detection. In particular Twitter[7] is a microblogging service that is gaining interests as a means for sharing real world events ranging from a user's personal status to news reports. Given its nature and volume, Twitter messages (or *tweets*) are now seen as a valuable source for real-time Web applications as trend detection and natural disaster detection. Sakaki et al. [126] use Twitter to detect a real-time event, such as, Earthquake. They propose a probabilistic spatio-temporal model for the target event that can find the center and the trajectory of the event location. More precisely, they model the probability of an event occurrence at time t using an exponential distribution in a homogeneous Poisson process and the location of the event is estimated using the Kalman filter and particle filter algorithms. Cataldi et al. [37] propose a novel topic detection technique that permits to retrieve in real-time the most emergent topics expressed by the twitter community.

In the medical domain, there has been a surge in detecting public health related tweets for Event-Based Epidemic Intelligence (e-EI). In general, health related tweets (e.g., user status updates or news) are commonly found in Twitter as, for example: (a) *"I have the mumps...am I alone?"*; (b) *"my baby girl has a Gastroenteritis so great!! Please do not give it to meee"*; (c) *"#Cholera breaks out in #Dadaab refugee camp in #Kenya http://t.co/...."*; (d) *"As many as 16 people have been found infected with Anthrax in Shahjadpur upazila of the Sirajganj district in Bangladesh"*. Such information can indicate the existence and magnitude of real-world health related events. Thus, Twitter can be considered as a collector of real-time information that could be used by health authorities as an additional information source for obtaining early warnings; thereby helping them to prevent and/or mitigate the public health threats. The focus has been on building classifiers for detecting self-reported illness [41, 137], syndromes [40] and ailments [118]. Moreover, recent work has focused on validating the timeliness of Twitter by correlating tweets with real-world outbreak statistics, such as, Influenza-like-Illness rates [118, 142] and detecting flu outbreaks [18, 43, 99]. The aforementioned works show the advantage of using Twitter for detecting real world events focusing on *common and seasonal* diseases, such as, influenza or dengue fever. Existing systems also rely on *particular countries* with a high density of Twitter users, such as, United Sates, United Kingdom or Brazil. To the best of my knowledge, none of these previous work

[7] http://twitter.com

have focused on an temporal analysis of Twitter data for *general diseases* that are not only seasonal, but also sporadic and that occur in low tweet-density areas like Kenya or Bangladesh.

Another important aspect dealing with event detection is the timing information related to an event occurred. *When did it begin?* or *How long will it last?* such questions related to *time* commonly arise when reading news about a particular event, e.g., wars, political movements, sports competitions, natural disasters or disease outbreaks. The answer to such questions can be regarded as the *temporal fact* about an event, which is defined as a time point for an instantaneous event, or a time span for an event with a known begin and end duration [80]. Temporal facts about an event can be captured by temporal expressions mentioned in documents, i.e., a *time point* and a *time period* as illustrated in the following sentences for two real-world events: the 2011 Arab Spring and the E.coli outbreak in Germany in 2011. (I) *The Arab Spring event was reported to begin in Tunisia on* ***January 11, 2011*** *when demonstrators protested chronic unemployment and police brutality.* (II) *An outbreak of severe illness is causing concern in Germany, where 3 women have died and 276 cases of hemolytic uremic syndrome have been reported since the* ***2nd week of May 2011****.* Knowing about the temporal facts of an event of interest is useful for both a journalist (in order to write a news story) or a news reader (in order to understand the news). Moreover, temporal facts are also leveraged in many application areas, e.g., answering *temporal questions*, and browsing or querying *temporal knowledge*. Existing work on extracting temporal facts for an event follows two main directions: 1) extract temporal expressions from unstructured text using time and event recognition algorithms [139, 146], and 2) harvest temporal knowledge from semi-structured contents like Wikipedia infoboxes [80]. Unfortunately, previous approaches in the first group have not considered the relevance of temporal expressions, while the latter method is only applicable for the limited number of events with infoboxes provided. A number of ranking models exploiting temporal information have been proposed, including [26, 53, 101, 107]. Li and Croft [101] incorporated time into language models, called time-based language models, by assigning a document prior using an exponential decay function of a document creation date. They focused on recency queries, such that the more recent documents obtain the higher probabilities of relevance. Diaz and Jones [53] used document creation dates to measure the distribution of retrieved documents and create the temporal profile of a query. They showed that the temporal profile together with the contents of retrieved documents can improve average precision for the query by using a set of different features for discriminating between temporal profiles. Berberich et al. [26] integrated temporal expressions into query-likelihood language modeling, which considers time uncertainty inherent to a query and documents, i.e., temporal expressions can refer to the same time interval even they are not exactly equal. Metzler et al. [107]

considered implicit temporal information needs. They proposed mining query logs and analyze query frequencies over time in order to identify time-sensitive queries.

There are also works that have focused on recency ranking [45, 59, 63, 85], while analyzing queries over time has been studied in [97, 133]. Kulkarni et al. [97] studied how users' information needs change over time, and Shokouhi [133] employed different time series analysis methods for detecting seasonal queries. For an entity-ranking task, Demartini et al. [50] analyzed news history (i.e., past related articles) for identifying relevant entities in current news articles. Kanhabua et al. [87] introduced the task of *ranking related news predictions* with the main goal of improving information access to predictions most relevant to a given news story. Another important aspect is presented by Strötgen et al. [141], where they studied the problem of identifying *top relevant temporal expressions* in documents.

2.4.4 Knowledge management applied for security issues

As organizations move more business processes online, protecting the confidentiality and privacy of the information used during these processes is essential. Because many automated processes rely on electronic documents that contain mission-critical, personal, and sensitive information, organizations must make significant investments to properly protect these documents [3]. The management of health care data has different security requirements. Among the others, the two primary requirements are: i) the communication and storage of private information should guarantee confidentiality and data integrity, ii) fine-grained access control policies are needed for different actors involved. Many access control models exploit the concept of data classification to protect critical resources, the Bell La Padula model [25] is a significant example of access control rules that are based on the security levels of the user-requestor and the resource-requested. Indeed, in the medical domain, many e-Health systems are designed to enforce fine-grain access control policies and the medical records are *a-priori* well structured to properly locate the different parts of the managed complex information. Several security problems occur when e-Health systems are applied in those contexts where new information systems have not been developed yet but "documental systems" are, in some way, introduced. In order to address security and access control for EHR systems, several solutions have been proposed [24, 31, 66]. Although these solutions use role based access control for security management none of these took into account the structure and the semantics of EHRs. A first step in this direction was made in [86]. This approach focuses on identifying and organizing EHRs by means of semantic interpretation of internal data so that access control policies can be specified to authorize EHRs portions data sharing.

This means that today documental systems improperly allow users to access a digitalized version of a medical record without having previously classified the critical parts. Any document is treated as a monolithic resource. The classification of critical elements of a not-structured documents is not easy at all; very often they contain ambiguous parts that are strongly related to the doctor activity; for example it is quite usual for a nurse to write in its portion some details of the diagnosis that is competence of the doctor or, sometimes, administrative person write down in the anagraphical parts also some information related to the patient's disease. Up to date, monolithic resources are protected at a course grain level and a permit/deny access rule can be applied to the whole document and not to the specific parts that constitute it.
A medical record contains patient's sensitive information; it is composed by several sections including patient's contact information, summary of doctor's visits, patient's diagnosis, medical and family history, list of prescriptions, health examinations, the therapy, etc. Moreover, the adoption of semantic approaches in policy management and in the security research fields is quite new in the literature; in [20] the authors propose an ontology-based policy translation approach that mimics the behavior of expert administrators, to translate high level network security policies into low level enforceable ones. In [49], a text mining method has been proposed to deal with large amounts of unstructured text data in homeland-security applications.

Chapter 3

The Framework for Semantic-based Knowledge Management and Document Processing

In several contexts as medical and juridical, knowledge management dealing with acquiring, maintaining, and accessing contents within data, can improve public and private services providers. Many difficulties and limitations arise when the information is not structured and contained in textual format as for example electronic documents, Web sources or paper document that have no support for machine-readable and processable activities. At this aim in this work is defined a framework for document processing and knowledge management. It analyzes texts and automatically extracts relevant information, concepts and complex relations, as events, organizing the not structured information. The framework takes in input textual information, transforms them and produce in output structured elements.

In this chapter it is presented the framework formalization as well as the definition of the data and the of the semantic methodology adopted in this work.

3.1 A framework for Knowledge Management and Document Processing

In several contexts as medical and juridical, knowledge management dealing with acquiring, maintaining, and accessing knowledge within data, can improve public and private

services providers. Many difficulties and limitations arise when the information is not structured and contained in textual format (for example electronic or paper document) i.e. without any support for machine-readable and processable activities. At this aim I have defined a framework for document processing, it analyzes texts and automatically extracts relevant information, concepts and complex relations organizing the not structured information. The framework takes in input documents belonging to a domain, transforms them and produce in output structured elements.

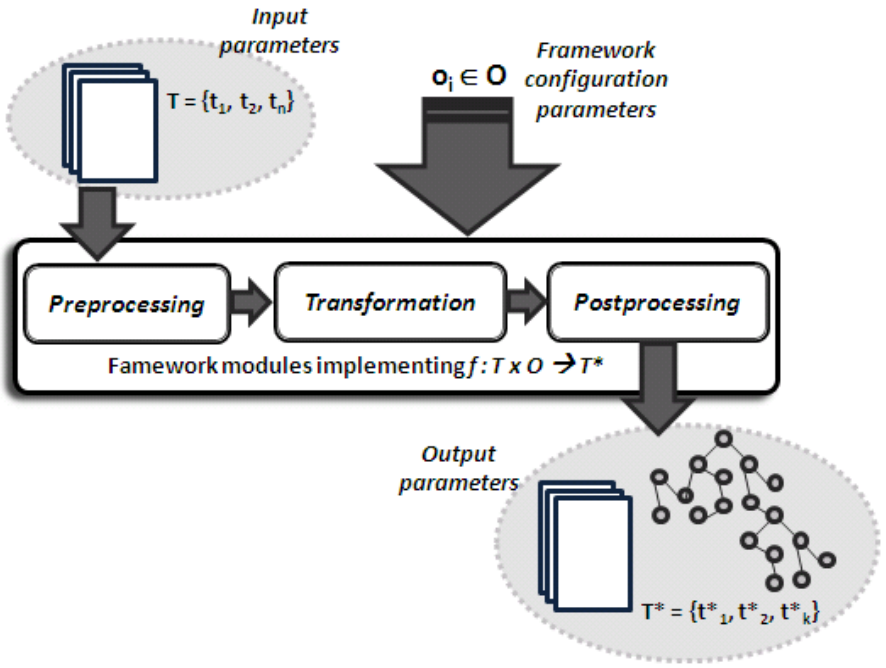

FIGURE 3.1: Document processing framework

The framework schema is depicted in Figure 3.1. It is composed of three main blocks: *(i)* preprocessing module for extracting textual elements from documents in input; *(ii)* transformation module for applying on textual elements a set of transformation rules, identified by the set of configuration parameters in input; *(iii)* postprocessing module to provide proper encoding of the textual elements according to different application scenarios. I have formalized the document processing framework as follows:

Definition 3.1 (**Document Processing Framework**). *The Document Processing Framework for a specific application domain D is a function*

$$f_D : T \times O \rightarrow T^*$$

where $T = \{t_1, t_2, ..., t_n\}$ *is the set of textual documents,* $O = \{o_1, o_2, ..., o_s\}$ *is the tuning set that defines the framework configurations and* $T^* = \{t_1^*, t_2^*, ..., t_k^*\}$ *is the outputted structured data (textual elements coded in a structured way as XML, RDF, etc.)*

Each module takes as input a subset of configuration parameters that select specific algorithms and techniques for documents transformation and eventually data inputs according to the context in which the framework has to be instantiated. It is possible to define the tuning set as follows:

Definition 3.2 (**Tuning Set**). *The Tuning Set O is defined as $O \subseteq A \times B \times C$ where*

$$A = \{\alpha_1, \alpha_2, ..., \alpha_h\}\ ,\ B = \{\beta_1, \beta_2, ..., \beta_l\}\ and C = \{\gamma_1, \gamma_2, ..., \gamma_r\}$$

are respectively the input parameters of the preprocessing, transformation and postprocessing modules.

In order to adopt the framework in a particular domain, it is necessary to perform a tuning phase by means of techniques, algorithms and input parameters selection. As depicted in Figure 3.1, the document processing framework takes as input the document set T and a set of configuration parameters that I have called *tuning parameters* and denoted with $O = \{o_1, o_2, ..., o_s\}$. Each instance of the framework identifies a specific tuning parameter $o_i \in O$ and vice versa. Multiple instance of the framework can together implement a system architecture. In the next chapter an instance of the framework will be described. It will be illustrated how the adoption of possible tuning parameters with selected tools produce several instances of the framework to produce a system architecture for knowledge management and document processing suited for the e-Health domain.

3.2 Data Models

In this section, I'll outline the data models used it this work for representing a document, an annotated document and a tweet. Those models will be adopted further in this work for event detection (Section 4.4).

3.2.1 Document and Annotated Document Models

A document collection is a set of outbreak reports composed of unstructured text documents: $C = \{d_1, \ldots, d_n\}$.

Definition 3.3 (**Document**). *A document d is defined as a bag-of-words or an unordered list of terms*

$$d = \{w_1, \ldots, w_k\}$$

where its publication date is denoted $PubTime(d)$.

Entity type	Term categories
Victims	Population, Age, Family, Animal, Food, Plant
Diseases	Medical Condition
Locations	City, ProvinceOrState, Country, Continent

TABLE 3.1: Named entities and their corresponding term categories.

For each document, it could be associated an *annotated document* $\hat{d}$.

A document collection is a set of reports composed of unstructured text documents: $C = \{d_1, \ldots, d_n\}$.

Definition 3.4 (**Annotated Document**). *An Annotated Document* $\hat{d}$ *is defined as:*

$$\hat{d} = \left(\hat{d_{ne}}, \hat{d}_t, \hat{d}_s\right)$$

where the component $\hat{d_{ne}}$ *represents a set of named entities* $\hat{d_{ne}} = \{ne_1, \ldots, ne_k\}$*; the component* $\hat{d}_t$ *is represented as a set of temporal expressions mentioned in d defined as* $\hat{d}_t = \{t_1, \ldots, t_h\}$*, denoted ContentTime*$(d)$ *and the component* $\hat{d}_s = (S, \leq_s)$ *is defined as a partially ordered set of sentences contained in d where* $S = \left\{s_1, \ldots, s_z \mid \bigcup_{j=1}^{z} s_j = d\right\}$ *and* $\forall i, j = 1, \ldots, z$ $s_i \leq_s s_j$ *means that the sentence* s_i *precedes* s_j *in the document d.*

In this work, the entities of interests are those relevant to the medical domain, i.e., diseases, victims, and locations. Table 3.1 presents the term categories of each type of named entities.

3.2.2 Tweet Model

A Twitter collection is defined as a set of tweets $T = \{tw_1, \ldots, tw_n\}$.

Definition 3.5 (**Tweet**). *A tweet tw is defined as:*

$$tw = (tw_{text}, tw_{loc}, tw_{time})$$

The contents of tweet tw_{text} *are represented as a bag-of-words or an unordered list of terms:* $tw_{text} = \{w_1, \ldots, w_k \mid \forall i = 1, \ldots, k\ w_i \in \mathcal{T}\}$ *where* $\mathcal{T}$ *is the set of allowed tokens that can be either a word, a hashtag (or user defined topic), or a URL. The time associated to a tweet, or* tw_{time}*, consists of its publication date PubTime*(tw) *and temporal expressions mentioned in the tweet contents. The* tw_{loc} *component represent the location where an event is occurred.*

In this work, the location information of a tweet (tw_{loc}) is the location of an outbreak event. Thus, location information is identified by choosing from the following sources

ordered by relevance: 1) text-contained location, 2) geolocation information (latitude and longitude), and 3) user's registered location.

3.3 A Methodology for Semantic Based Resource Characterization

In order to properly locate and characterize resources made of text sections, it is necessary to apply semantic text processing techniques on available data [7]. Semantic processing of documents is based on the knowledge and interpretation given by the document author that may not be the same of the reader.

The comprehension of a particular concept within a specialized domain, as for example the medical one, requires information about the properties characterizing it, as well as the ability to identify the set of entities the concept refers to. A text, in fact, is the product of a communicative act resulting from a process of collaboration between an author and a reader. The former uses language signs to codify meanings, the latter decodes these signs and interprets their meaning by exploiting the knowledge of:

1. the *infra-textual* context, consisting in relationships at a morphological, syntactic and semantic level;
2. the *extra-textual* context and, more in general, the *encyclopedic knowledge* involving the domain of interest.

Starting from these points, the activity of knowledge extraction from texts includes different kinds of text analysis methodologies, aiming at recreating the model of the domain the text pertain to. In the next subsection, it is illustrated the process of extracting information from documents. To better explain the stages of the methodology, I will use a fragment of a psychiatric medical record as a running example. It states that, at the entrance of the hospital, a patient results quiet and cooperative, calm in the maxilla-facial expression:

```
Diagnosi di entrata la paz. e' tranquilla
e collaborante, serena nell' espr. maxillofacciale
```

It is worth to note that the example refers to a medical record in Italian, nevertheless the proposed approach is general enough to be applicable to other languages as well as other domains.

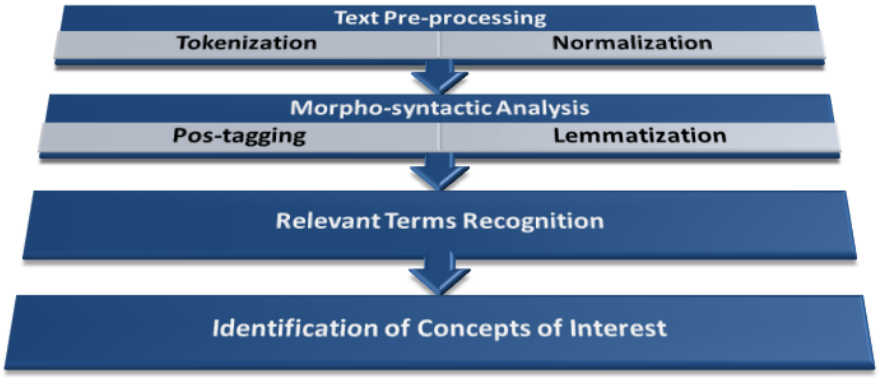

FIGURE 3.2: Stages of the methodology for concepts identification

3.3.1 Stages of the Methodology

Term-extraction is a fundamental activity in the automatic document processing and derivation of knowledge from texts.

Terms serve to convey the fundamental concepts of a specific knowledge domain: they have their realization within texts and their relationships constitute the semantic frame of the documents and of the domain itself. The main goal is to find a series of relevant and peculiar terms in order to detect the set of concepts that allow the resource identification.

In order to extract relevant terms from text, it is used an hybrid method that combines *linguistic and statistical techniques*: it is employed a *linguistic filter* in order to extract a set of candidate terms and then use a *statistical method* to assign a value to each candidate term. In particular, linguistic filters are applied on the words, like as *part-of-speech tagger* (aiming at extracting the categories of interest, such as nouns and verbs), and *lemmatization* (that restore words to a dictionary form).

Statistical methods are based on the analysis of word occurrences within texts, in order to measure the "strength" or "weight" of a candidate term. As a matter of fact, not all words are equally useful to describe documents: some words are semantically more relevant than others, and among these words there are lexical items weighting more than others.

In order to extract relevant terms from a medical record, several steps are required; these are described in details in the following sections and illustrated in Figure 3.2.

Text Preprocessing. This stage aims at extracting processable plain text from the input documents, by detecting units of lexical elements that can be processed in the next stages. It implements text tokenization and text normalization procedures.

Text tokenization consists in segmentation of sentences into tokens, minimal units of analysis, which constitute simple or complex lexical items, including compounds, abbreviations, acronyms and alphanumeric expressions.
Text tokenization requires, various sub-steps, as: *grapheme analysis*, to define the set of alphabetical signs used within the text collection, in order to verify possible mistakes as, for example, typing errors, misprints or format conversion; *disambiguation of punctuation marks*, aiming at token separation; *separation of continuous strings* (i.e. strings that are not separated by blank spaces) to be considered as independent tokens: for example, two terms separated by the character " ' "; and *identification of separated strings* (i.e. strings that are separated by blank spaces) to be considered as complex tokens and, therefore single units of analysis.
This segmentation can be performed by means of special tools, defined *tokenizers*, including *glossaries* with well-known expressions to be regarded as medical domain tokens and *mini-grammars* containing heuristic rules regulating token combinations. The combined use of glossaries and mini-grammars ensures high level of accuracy, even in presence of texts with acronyms or abbreviations that can increase the mistakes rate. Considering the running example, the output of text tokenization is:

```
Diagnosi//di//entrata//la//paz.//
e'//tranquilla//e//collaborante,//
serena//nell//'//espr.// maxillofacciale//
```

Text normalization takes variations of the same lexical expression back in a unique way; for example, *(i)* words that assume different meaning if are written in small or capital letter, *(ii)* compounds and prefixed words that can be (or not) separated by a hyphen, *(iii)* dates that can be written in different ways ("1 Gennaio 1948" or "01/01/48"), *(iv)* acronyms and abbreviations ("USA" or "U.S.A.", "pag" or "pg"), etc.
The transformation of capital letters into small letters, is a not trivial operation: for example, a capital letter helps in identifying the beginning of a sentence and differentiating a common noun (like the flower "rosa") from a proper name (such as "Rosa") or even to recognize the distinction between an acronym (e.g."USA") and a verb (e.g. "usa", 3rd sing. pers. of the Italian infinitive "usare"). The output of this phase is, for the running example:

```
Diagnosi//di//entrata//la//
paziente//e'//tranquilla//e//collaborante,//
serena//nell//'//espressione//maxillo-facciale//
```

Morpho-syntactic analysis. The main goal of this stage is the extraction of word categories, both in simple and complex forms. This leads to obtain a list of candidate terms on which relevant information extraction can be performed.

Part-of-speech (POS) tagging consists of the assignment of a grammatical category (noun, verb, adjective, adverb, etc.) to each lexical unit identified within the text collection.

Morphological information about the words provides a first semantic distinction among the analyzed words. The words can be categorized in: *content words* and *functional words*. Content words represent nouns, verbs, adjectives and adverbs. In general, nouns indicates people, things and places; verbs denote actions, states, conditions and processes; adjectives indicate properties or qualities of the noun they refer to; adverbs, instead, represent modifiers of other classes (place, time, manner, etc.). Functional words are made of articles, prepositions and conjunctions; they are very common in the text.

Automatic POS tagging involves the assignment of the correct category to each word encountered within a text. But, given a sequence of words, each word can be tagged with different categories [67].

As already stated, the *word-category disambiguation* involves two kinds of problems: *i)* finding the POS tag or all the possible tags for each lexical item; *ii)* choosing, among all the possible tags, the correct one. Here the vocabulary of the documents of interest is compared with an external lexical resource, whereas the procedure of disambiguation is carried out through the analysis of the words in their contexts. In this sense, an effective help comes from the *Key-Word In Context* (KWIC) analysis, a systematic study of the local context where the various occurrences of a lexical item appear. For each concept it is possible to locate its occurrences in the text and its co-text (i.e. the textual parts before and after it). The analysis of the co-text, then, allows detecting the role of the words in the phrase, in order to disambiguate their grammar category. The ambiguous form is then firstly associated to the set of possible POS tags, and then disambiguated by resorting to the KWIC analysis. The set of rules defining the possible combinations of sequences of tags, proper of the language, enables the detection of the correct word category. Consider, in the reported example, the ambiguity associated to the Italian word “entrata”: it can be a noun (“entry”) or a verb (“enter”). This ambiguity can be solved by analyzing the categories of the preceding words: rules derived by syntax of Italian language state that, if the word is preceded by an article or a preposition it is a noun, while if it is preceded by a noun it is a verb. Then, applying KWIC analysis it derives that “entrata” is a verb.

Further morphological specifications, such as inflectional information[1], are then associated to each word.

The output of this stage, for the running example, is:

[1]Inflection is the way language handles grammatical relations and relational categories such as gender (masculine/feminine) and number (singular/plural) for nouns; tense, mood, person and voice for verbs.

```
Diagnosi          NOUN
di                PRE
entrata            NOUN
la                ART
paziente          NOUN
tranquilla        ADJ
e                  CON
collaborante       NOUN
,                 PUN
serena            ADJ
nell'             ARTPRE
espressione        NOUN
maxillo-facciale  NOUN
```

Note that, were used the following conventions: (ART) = article; (ADJ) = adjective; (ADV) = adverb; (CON) = conjunction; (NOUN) = noun; (PN) = pronoun; (PRE) = preposition; (VERB) = verb; (ARTPRE)= article + preposition.

Lemmatization is performed on the list of tagged terms, in order to reduce all the inflected forms to the respective lemma, or citation form, coinciding with the singular male/female form for nouns, the singular male form for adjectives and the infinitive form for verbs. The output of this stage, for the running example is:

```
Diagnosi          NOUN    diagnosi
di                PRE    di
entrata            NOUN    entrata
la                ART    il
paziente          NOUN    paziente
tranquilla        ADJ    tranquillo
e                  CON    e
collaborante      NOUN   collaborante
,                 PUN    ,
serena            ADJ    sereno
nell'             ARTPRE nel
espressione        NOUN    espressione
maxillo-facciale  NOUN    maxillo-facciale
```

Note that many terms are already present in canonical form, and for this reason, in this phase, they are not converted; while the other terms, as the adjective "tranquilla" or the preposition "nell" are respectively transformed in "tranquillo" and "nel".

Relevant Terms Recognition. The goal of the methodology is the identification of the relevant terms, useful to characterize the sections of interest [10]. In fact, as state above, not all words are equally useful to describe resources: some words are semantically more relevant than others, and among these words there are lexical items weighting more than other. In this approach, the semantic relevance is evaluated by the assignment of the *tf-idf* index (*Term Frequency - Inverse Document Frequency* [128]), computed on

the corpus vocabulary and on the base of the term frequency and the term distribution within the corpus. *tf-idf* index, in fact, takes into account:

- the *term frequency* (*tf*), corresponding to the number of times a given term occurs in the resource: the more a term occurs in the same section, the more it is representative of its contents. Frequent terms are then supposed to be more important. This method is used in systems to rank terms candidates generated by linguistic methods [44].
- the *inverse document frequency (idf)*, concerning the term distribution within all the sections of the medical records: it relies on the principle that term importance is inversely proportional to the number of documents from the corpus where the given term occurs. Thus, the more resources contain that given term, the less discriminating it is.

Therefore, *tf-idf* enables the extraction of the most discriminating lexical items because they are frequent and concentrated on few documents. This statement is summarized in the following ratio:

$$W_{t,d} = f_{t,d} * log(N/D_t)$$

where $W_{t,d}$ is the evaluated weight of term t in resource d; $f_{t,d}$ is the frequency of term t in the resource d; N is the total number of occurrences within the examined corpus; D_t is the number of resources containing the term t.

For the running example, this phase produces the following information:

```
diagnosi          5     *
entrata           1,5
paziente          4     *
tranquillo        2,8
collaborante      3,1   *
sereno            2,5
espressione       3,8   *
maxillo-facciale  7     *
```

This information enables the selection of relevant concepts, filtering all terms that have a *tf-idf* value under an established threshold. It was used, as threshold, the value 3: all terms whose *tf-idf* is over this threshold will be considered relevant. In the example, the relevant terms are marked with an asterisk.

Identification of Concepts of Interest. Once relevant terms, belonging to the used medical sub-domain, are detected, I proceed to cluster them in **synset**, *"a group of data elements that are considered semantically equivalent for the purposes of information retrieval"*, in order to associate the semantic concept that every cluster of terms refers to.
In this way it is possible referring to a concept independently from the particular term used to indicate it. Examples of the use of concepts, codified as synsets, for identifying sections of text are shown in [11] for the medical domain, and in [9] for the legal domain.
To group the term two external resources areused: the medical ontology given by Unified Medical Language System (UMLS) [102] and "Mesh"[2], a thesaurus of medical terms.
The adoption of specialized external resources has a duplicate purpose:

- **Endogenous:** Inside the documental base, the same concepts can be referred by different terms.
- **Exogeneous:** An user, that can query the documental base with an interrogation written in natural language, can use, for indicate a certain concept, a term that is different from those used in the documental base, and then do not appear in it.

Every concept is identified by a synset (i.e. the set of synonyms), I associate each term extracted from the medical record to a synset by a unique label that represents a witness for the given synset.
This stage associates the synset, i.e. the proper concept, to each selected term of the running example, as showed in the Table 3.2.

Term	**Synset**	**Label**
diagnosi	parere, prognosi, responso, valutazione, analisi	Diagnosi
paziente	ammalato, degente, malato	Paziente
espressione	manifestazione, segno, smorfia, viso, sintomo	Sintomo
maxillo-facciale	maxillo-facciale	Maxillo-Facciale

TABLE 3.2: Association of synsets with concepts

The table shows that for each relevant term, extracted on the basis of its grammatical category and *tf-idf* value, it is associated a synset: a set of terms referring the same concept. This list of terms is built by exploiting the relation codified in the external domain resources: UMLS and "Mesh". In this example the concepts associated to the extracted terms were: "Diagnosi" (diagnosis) , "Paziente" (patient), "Sintomo" (Symptom) and "Maxillo-Facciale" (maxillofacial).

[2]Medical Subject Headings of National Library of Medicine. http://www.nlm.nih.gov/mesh/

Chapter 4

The Framework Instance: an Architecture for the e-Health

In the health domain, the information availability coming from different sources can improve the health services quality. For example, when a doctor has to deal with a diagnosis task, it could result very difficult to determine the patient's disease because of few and common nature of symptoms. Moreover a many diseases have several symptoms in common and knowing the place of origin of the disease can improve the diagnosis task and also the treatment assignment. For example, a disease that is common and frequent in a country can occur in a place in which is unusual and rare leading to a possible diagnosis mistake. In this scenario, the doctor's work and the patient's care could be improved through the use of methodologies and technologies that facilitate the actors of the medical domain in accessing health information from various sources such as, for example, medical records of hospital departments belonging to different countries and information about outbreaks currently occurring world wide. To this aim, the medical data must be properly organized and information from the Web must be conveniently filtered and monitored in order to automatically detect events related to infectious diseases currently occurring. In this chapter is provided an architecture as multiple document processing framework instances. This architecture exploits semantic-based methodology in order to process data and extract information, as concepts or complex relations, in order to implement several functionalities described in the following. The architecture provides the answers to the research questions defined in the Introduction (Section 1.1):

Question 1: Is it possible to automate processes for data classification of paper documents? And, how?

Question 2: Can be the access policies applied automatically for fine-grain protection of information?

Question 3: How external sources can be exploited for complementing traditional information?

To answer those questions, the framework for knowledge management and document processing has been instantiated, applied and tested (as will be described in the Chapter 5) in the medical domain.

4.1 An Architecture for the e-Health Knowledge Management and Medical Records Processing

In this work it has been defined an architecture for heterogeneous and multi-language data management, in order to support the actors in the medical domain to accessing and retrieving useful information. In particular, this architecture supports the user in the medical record composition allowing to:

- Organize previously scanned medical records according to the field of diagnosis (hospital departments such as surgery, cardiology, etc..) [16];
- Structuring the medical records and identify the sections to associate automatically access policies [11–15];
- Manage information from external sources in order to identify outbreaks of epidemics [88, 89].

The whole architecture is depicted in figure 4.1. The main blocks of the architecture are described in the following sections. In particular, in Section 4.2 will be described how to use the framework for medical records classification; in Section 4.3 will be described how to adopt the framework for automatically associating fine grain access polices to medical records; and in Section 4.4 will be described how to adopt the framework for event detection from Web sources.

4.2 Adopting the Framework for Medical Records Classification

In this section it is described the architecture for document classification. It will be described how textual information are processed and then automatically organized by means of the clustering ensemble method. At this aim three Vector Space Models (VSM)

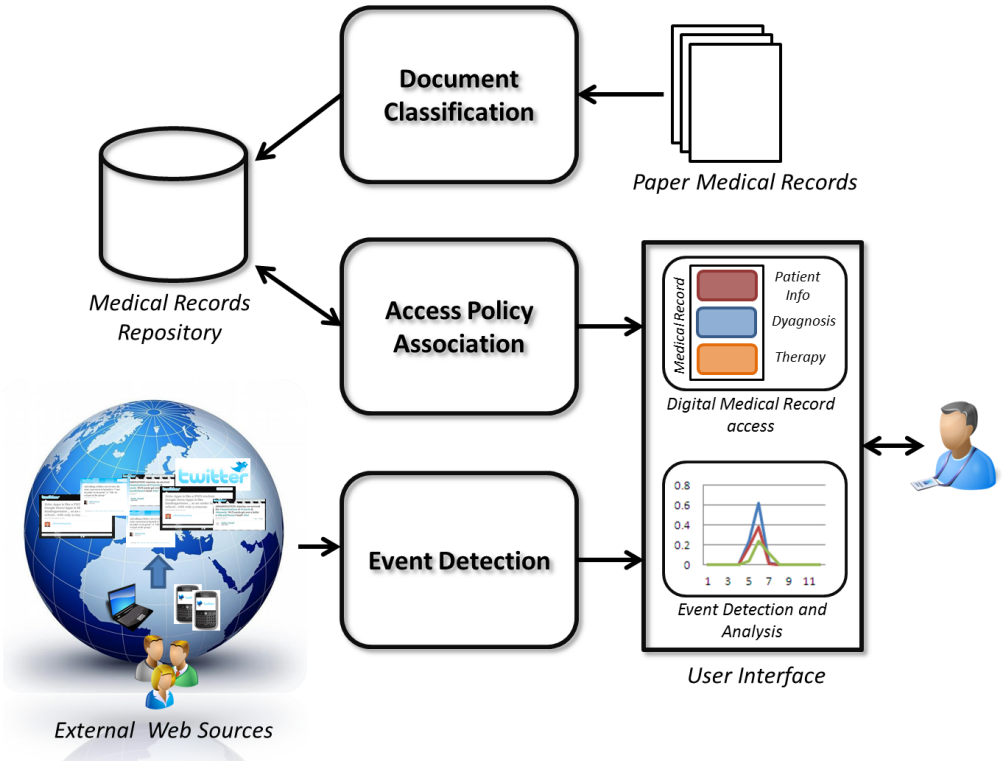

FIGURE 4.1: The Architecture for Knowledge Management and Document Processing in the e-Health

for document representation have been adopted. Those include syntactic and semantic aspects based respectively on frequencies of terms, lemmas and concepts.

Clustering ensemble (or clustering aggregation) is an alternative approach that combines different clustering results in order to improve the quality of clustering results. In general, a clustering ensemble method is composed by two steps: generation and consensus. The generation step consists on the production of the set of clusterings obtained with different clustering algorithms or the same algorithm with different parameters initialization. The consensus step represents the main challenge in the clustering ensemble algorithm. In this step, a function of consensus that take into account the results of single clustering algorithms is defined. The result of this step is the final data partition (or consensus partition). The main approach for defining the consensus function is objects co-occurrence where it is investigated how many times an object belongs to one cluster or how many times two objects are associated to the same cluster. The methods based on this approach are known as Co-association Matrix and Relabeling and Voting. The median partition approach is based on the optimization problem aiming to find the median partition with respect to the cluster ensemble. This approach is followed by Kernel methods and Non-Negative Matrix Factorization. In literature there are several works that address the problem of document categorization by means of clustering ensemble techniques [58, 71, 73, 132]. In [145] is presented a good analysis

of the existing techniques of clustering ensemble method. In [71] the authors propose a compound ensemble clustering algorithm combining the statistic information of the data with the sense information from WordNet. Their ensemble clustering framework combines the k-means clustering solutions obtained from the semantic similarity of the document (semantic binary model) with those obtained on frequency similarities (nouns frequency model) and produce the final result exploiting the co-association matrix.

In Figure 4.2 is presented the framework instance for document classification.

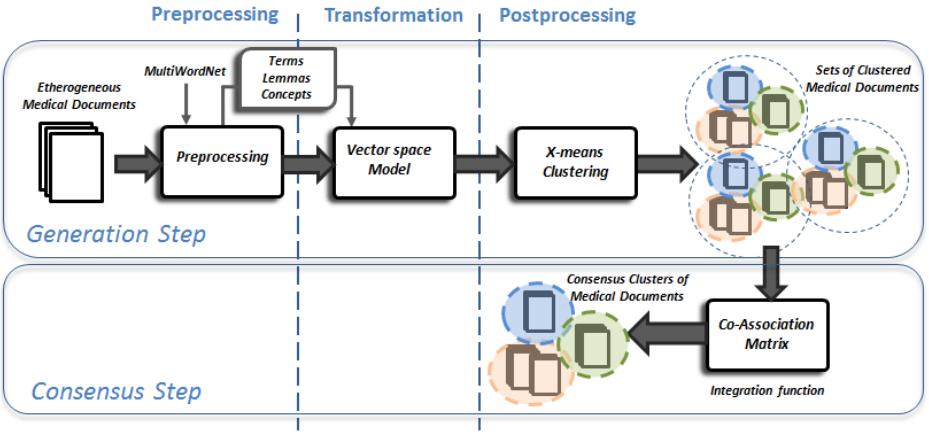

FIGURE 4.2: Framework instance for document classification

In this work, in order to represent the document collection in the vector space model, it is necessary to extract the feature by means of Natural Language Processing (NLP) steps. By using a mix of lexical and statistic procedures, reported in the following sections, were extracted a set of terms from the document corpus, and therefore, the set of synonyms corresponding to them. At this aim it was adopted the semantic methodology described in Section 3.3 for the automatic extraction of concepts of interest.

The implemented set of procedures aiming at extracting terms (Criterion I), the corresponding lemmas (Criterion II) and the associated concepts (Criterion III) from the input documents are described in the following.

Extracting Terms (I Criterion): Starting from the input documents, by using *Text Tokenization* procedures, text is arranged into tokens, sequences of characters delimited by separators. Applying *Text Normalization* procedures,variations of the same lexical expression are reported in a unique way.
Tokenization and Normalization procedures perform a first grouping of the extracted text, introducing a partitioning scheme that establishes an equivalence class on terms. At this point I built the doc-features matrix, having a column for each term in the terms list, which contains the evaluation, for each document, of the *tf-idf* value for

every terms in the list. The *tf-idf* values are computed taking into account both the number of occurrences of each term for every documents and the terms distribution in the whole document corpus. This matrix is considered as input for the clustering algorithm according to the I Criterion.

Extracting Lemmas (II Criterion): In order to obtain the lemmas starting from the list of relevant text, procedures of *Part-Of-Speech*(POS) *Tagging* and *Lemmatization* have been applied. These procedures aim at enriching the text with syntactical aspects, and performing a second type of grouping of the words, on the basis of reduction of terms in a basic form, independently from the conjugations or declinations in which they appear. *Part-Of-Speech (POS) Tagging* consists in the assignment of a grammatical category to each lexical unit, in order to distinguish the content words representing noun, verb, adjective and adverb from the functional words, made of articles, prepositions and conjunctions, denoting not useful information.
Text Lemmatization is performed in order to reduce all the inflected forms to the respective lemma, or citation form. Lemmatization introduces a second partitioning scheme on the set of extracted terms, establishing a new equivalence class on it.
It was built a doc-features matrix, having a column for each lemma in the list, which contains, for each document, the *tf-idf* value of each lemma comparing in it. This value is computed considering the sum of the number of occurrences of each term that can be taken back to the same lemma appearing in the document. The lemma based doc-features matrix is considered as input for the clustering algorithm according to the II Criterion.

Extracting Concepts (III Criterion): In order to identify concepts, not all words are equally useful. Some of them are semantically more relevant than others, and among these words there are lexical items weighting more than others. In order to "weight" the importance of a term in a document, it was adopted also in this case the *tf-idf* index.
Having the list of relevant terms, concepts are detected by relevant token sets that are semantically equivalent (synonyms, arranged in sets named synset). In order to determine the synonym relation among terms, it was exploited as external resource WordNet [109] a thesaurus codifying the relationship of synonymy among terms.
The number of occurrence of a concept in a document is given by the sum of the number of occurrences of all terms in its synonym list that appear in the document. The concept based doc-features matrix contains, for each document, the *tf-idf* of every concepts comparing in it. The *tf-idf* values of such matrix is then evaluated on the basis of the sum of the number of occurrences of each terms that is synonym of the input terms,

i.e. that is included in the synonym list. The concept based doc-features matrix is considered as input for the clustering algorithm according to the III Criterium.

Once the documents are represented by vector space models, those are then used for the clustering algorithm.

In this section is proposed the combination of a set of clusters for exploiting the different information levels provided by both syntactic and semantic features. As base cluster, it was used the X-means algorithm [120] that can be considered as an evolution of the standard K-means approach.

4.2.1 Clustering ensemble steps

The clustering ensemble method used [19] is shown in Figure 4.3. It is composed by the following steps:

1. It is considered the initial document matrix A for each criteria *Terms*, *Lemmas* and *Concepts*. A is a $n \times m$ matrix where n is the number of documents and m depends on the criteria selected and it could represent the number of *terms*, *lemmas* or *concepts*.

2. Are generated $C_k, k = 1, 2, \ldots, L$ partitions of A by using the X-means algorithm. Each partition have a random number of clusters depending on the initial seed chosen.

3. It is defined a co-association matrix for each partition: $M^k = m^k_{i,j}$, of dimension $n \times n$, where n is the number of documents and $k = 1, 2 \ldots, L$. The elements of the matrices M^k are calculated as:

$$m^k_{i,j} = \begin{cases} 1 & a_i = a_j \text{ (i.e. in the same cluster)}, \\ 0 & \text{otherwise.} \end{cases}$$

4. It is defined a co-association final matrix obtained as $M = 1/L * \sum_{k=1}^{L} M^k$.

5. It is selected a threshold σ that maximizes the adopted performance indexes [98], and then it is used an inverse function, denoted as *Clustering Evaluation* in the Figure 4.3, in order to obtain the final documents partition C from M and σ.

6. All the obtained results are compared by using the Rand Index, the Normal Mutual Information (NMI) index [98] and the number of generated clusters.

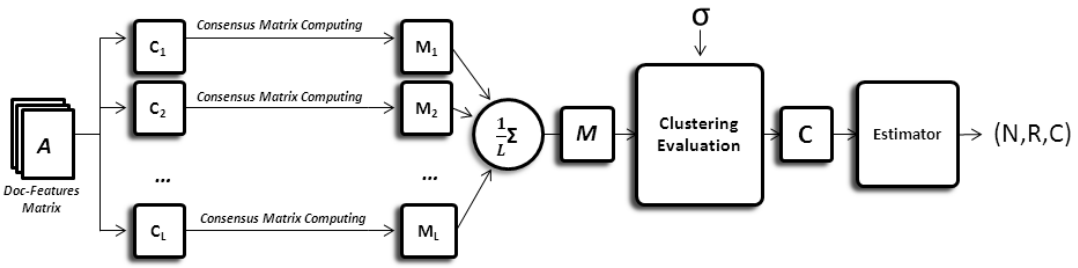

FIGURE 4.3: Generation and evaluation of the proposed clustering solution

4.3 Adopting the Framework in the Security Domain: Fine-grain Access Policies

Organizing not structured text in order to obtain the same information content in a semantically structured fashion is a challenging research field that can be applied in different contexts. For example, a possible scenario is represented by structuring information using a semantic approach (with techniques for knowledge extraction and representation) in order to develop a semantic search engine that enables the access to information contents (semantic) of a not structured document set. Moreover a semantic approach can be used on unstructured information in order to detect sensible information and for enforcing fine grained access control on these.

In the medical domain, the management of health care data has different security requirements. Among the others, two primary requirements are: i) the communication and storage of private information should guarantee confidentiality and data integrity, ii) fine-grained access control policies are needed for different actors.

As illustrated in Figure 4.4, a medical record is a structured data made of different parts each of these can be read and/or modified by different actors. Many of this data is private and can be viewed only by patients and their doctors, other parts are registry or administrative information and should be viewed only by administrators of the hospital. Analyzing the security requirements associated to such data and the result has led to state that an access control model that strongly takes in consideration the attributes of the resources to protect and the actor role should be enforced in e-health systems.

These scenarios have motivated the proposal of the framework instances. In this section I'll show how to adopt the framework for sensible resources detection in medical records in order to enforce fine-grain protection.

In order to properly locate and characterize text sections, was applied the semantic text processing methodology [7] described in Section 3.3. The comprehension of a particular concept within a specialized domain, as the medical one, requires information about

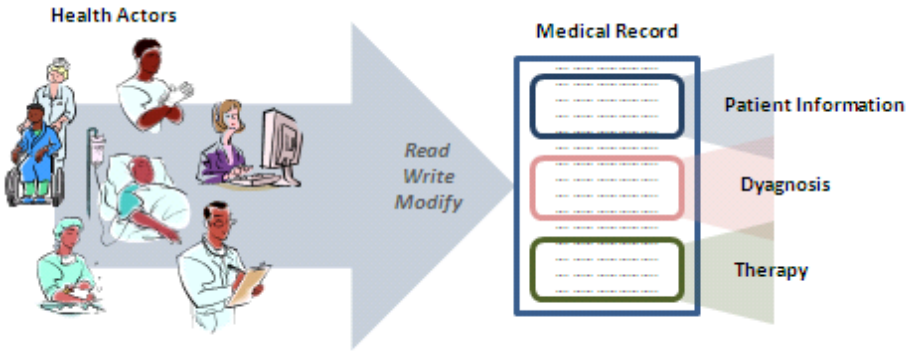

FIGURE 4.4: Actors of health domain accessing to medical records and their sections

the properties characterizing it, as well as the ability to identify the set of entities that the concepts refer to. At this aim the preprocessing, transformation and postprocessing modules should respectively: *(i)* breaking up a stream of text into a list of words and phrases, marking up the tokens as corresponding to a particular part of speech; *(ii)* filtering the token list obtaining the most relevant ones in order to build concepts; *(iii)* identifying the text macro-structures (sections). At this aim I have instantiated the framework as depicted in Figure 4.5 and described as follows.

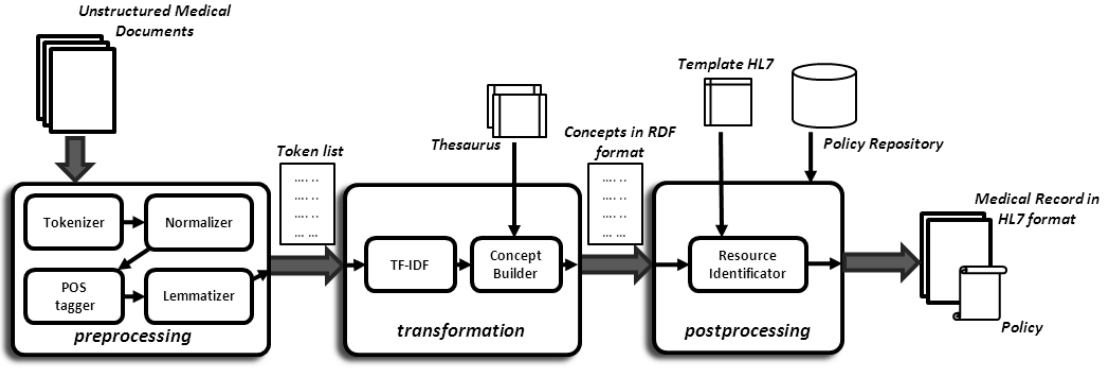

FIGURE 4.5: Framework instance for securing documents and sections

In order to process documents belonging to the E-Heath domain, the document processing framework has been instantiated by means of input parameters o_i selection and corresponding techniques. In this way specific tools have been selected in order to implement each module. The system accepts in input the corpus (made of unstructured medical records) and performs the formalization activities in order to structure it. In this way the resources to protect, representing the objects of the security rules are easily located.

In particular, in the *Preprocessing* module, all procedures are implemented by choosing a suitable tool for text analysis: TaLTaC^2 [131] (in Italian language). In the *Trasformation* module the procedure responsible for *tf-idf* calculation is computed again by

TaLTaC[2]. The *Concepts Builder* component is implemented by means of an innovative software designed by our research group. It takes as input a list of relevant words (those having higher *tf-idf* value) and, exploiting a domain thesaurus[1] for semantic relations identification, clusterize them in concepts.

The resources identification of the *Postprocessing* module uses the classification procedure offered by the KNIME[2] workflow tool.

In order to process the input text and produce a list of words, the **preprocessing module** implements in sequence: *Text Tokenization*, *Text Normalization*, *Part-Of-Speech (POS) Tagging* and *Lemmatization* procedures. The main goal of these procedures is the extraction of relevant terms that are used to recognize concepts in the text. Text Tokenization and Text Normalization procedures perform a first grouping of the extracted terms, introducing a partitioning scheme that establishes an equivalence class on terms.

Text Tokenization segments text into minimal units of analysis; *Text Normalization* takes variations of the same lexical expression back in a unique way. In particular Text tokenization includes many sub-steps as *grapheme analysis*, to define the set of alphabetical signs used within the text collection in order to verify possible mistakes, as typing errors, misprints or format conversion; *disambiguation of punctuation marks*, aiming at token separation; *separation of continuous strings* i.e. strings that are not separated by blank spaces to be considered as independent tokens: for example, in the italian string "l'anestesista" there are two independent tokens ("l' " + "anestesista"); *identification of separated strings* i.e. strings that are separated by blank spaces to be considered as complex tokens and, therefore, single units of analysis.

This segmentation can be performed by means of special tools, defined *tokenizers*, including *glossaries* with well-known expressions to be regarded as medical domain tokens and *mini-grammars* containing heuristic rules regulating token combinations. The combined use of glossaries and mini-grammars ensures high level of accuracy, even in presence of texts rich of acronyms or abbreviations, as the medical one, that can increase the mistakes rate.

Text normalization procedures take variations of the same lexical expression that should be reported in a unique way, such as words that assume different meaning if are written in small or capital letter; compounds and prefixed words that can be (or not) separated by a hyphen; dates that can be written in different ways; acronyms and abbreviations as "CAP" or "C.A.P.". This phase is also responsible for the transformation of capital letter that, for example, helps in distinguish a common noun used at beginning of a

[1]The Medical Subject Headings comprise the National Library of Medicine's - www.nlm.nih.gov/mesh/

[2]http://www.knime.org/

sentence from a proper name (for example to distinguish between the acronym "USA" and the Italian verb "usa" standing for "use").

The procedures of Part-Of-Speech (POS) Tagging and Lemmatization aim at enriching the text of meta-information about syntactical aspects associated to the extracted tokens, aiming at performing a second type of grouping of the words, on the basis of reduction of terms in a basic form, independently from the conjugations or declinations in which they appear.

These operations are performed in order to detect relevant words, by filtering out the text of grammatical words not carrying useful information.

Part-Of-Speech (POS) Tagging consists in the assignment of a grammatical category (noun, verb, adjective, adverb, etc.) to each lexical unit identified within the text collection. Morphological information about the words provides a first semantic distinction among the analyzed words. The words can be categorized in: *content words* and *functional words*. Content words represent nouns, verbs, adjectives and adverbs. In general, nouns indicates people, things and places; verbs denote actions, states, conditions and processes; adjectives indicate properties or qualities of the noun they refer to; adverbs, instead, represent modifiers of other classes (place, time, manner, etc.). Functional words are made of articles, prepositions and conjunctions; they are very common in the text.

Automatic POS-tagging involves the assignment of the correct category to each word encountered within a text. But, given a sequence of words, each word can be tagged with different categories [67].

The *word-category disambiguation* involves two kinds of problems: *i)* finding the POS-tag or all the possible tags for each lexical item; *ii)* choosing, among all the possible tags, the correct one. Here the vocabulary of the documents of interest is compared with an external lexical resource, whereas the procedure of disambiguation is carried out through the analysis of the words in their contexts.

Text Lemmatization is performed in order to reduce all the inflected forms to the respective lemma, or citation form, coinciding with the singular male/female form for nouns, the singular male form for adjectives and the infinitive form for verbs. Lemmatization introduces a second partitioning scheme on the set of extracted terms, establishing a new equivalence class on it.

All these procedures are language dependent, consisting of several sub-steps, and are implemented by using the state of the art NLP modules [105].

At this point, a list of tokens is obtained from the raw data. In order to identify concepts, not all words are equally useful: some of them are semantically more relevant than others, and among these words there are lexical items weighting more than other. The **transformation module** aims at filtering the token list in order to obtain a reduced list, containing only the relevant tokens. To do that, there are several techniques in literature that "weight" the importance of a term in a document, based on the statistics of occurrence of the term. *Tf-idf* index (*Term Frequency - Inverse Document Frequency*) [44] is actually the most popular measure used to evaluate terms semantic relevance.

Having the list of relevant terms, concepts are detected by relevant token sets that are semantically equivalent (synonyms, arranged in sets named synset). In order to determine the synonym relation among terms, it is possible to use statistic-based techniques of unsupervised learning, as clustering, or external resources like thesaurus (an example is wordnet). At this point it is possible to codify concepts by means of ontology data models (RDF, OWL, etc..).

Once the set of textual concepts is identified, the **postprocessing module** performs resource recognition that implies textual macro-structures identification. It consists of a classification task, exploiting the presence or the absence of concepts identified in the previous section. In literature the classification process is a statistic-based technique based on supervised learning. There are several implementations of classifiers as, for example, Naive Bayes [2], Decision Tree [124], K-Nearest Neighbor [156]. The textual macro-structures are the sensible resource to protect and represent objects of the security roles. Once these are identified, it is possible to apply a fine grain access control policy allowing users to perform only authorized actions on the resources according to their role. Furthermore it is possible to codify the identified resources in a structured fashion by means of available data models (as for example XML, HL7, etc.).

4.4 Adopting the Framework for Knowledge Management: Event Extraction from Twitter

Health related tweets (e.g., user status updates or news) are commonly found in Twitter as, for example: (a) *"I have the mumps...am I alone?"*; (b) *"my baby girl has a Gastroenteritis so great!! Please do not give it to meee"*; (c) *"#Cholera breaks out in #Dadaab refugee camp in #Kenya http://t.co/...."*; (d) *"As many as 16 people have been found infected with Anthrax in Shahjadpur upazila of the Sirajganj district in Bangladesh"*. Such information can indicate the existence and magnitude of real-world health related events. Thus, Twitter can be considered as a collector of real-time information that

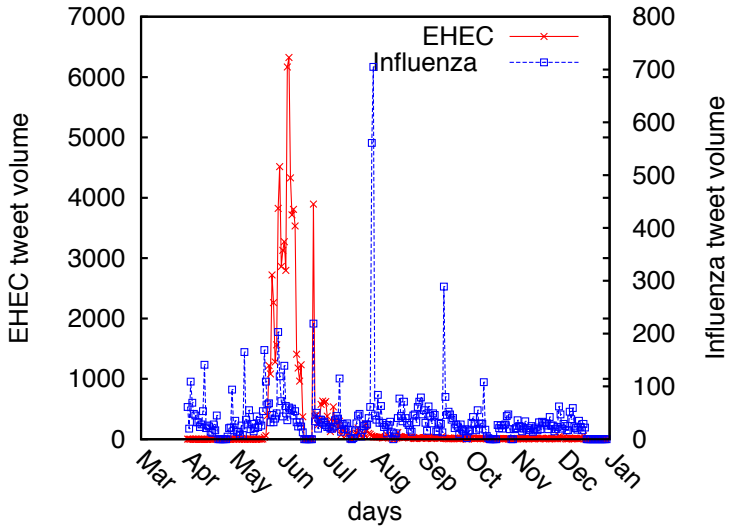

FIGURE 4.6: Distributions over time of tweets related to two outbreak events: (1) *EHEC* outbreak in Germany in May 2011 and (2) *avian influenza* in Cambodia in August 2011.

could be used by health authorities as an additional information source for obtaining early warnings; thereby helping them to prevent and/or mitigate the public health threats.

Recent work has focused on validating the timeliness of Twitter by correlating tweets with real-world outbreak statistics, such as, *Influenza*-like-Illness rates [118] and detecting flu outbreaks [18, 43, 99]. Figure 4.6 illustrates the distributions over time of tweets containing keywords ehec and avian influenza, where the highest peaks in both time series correspond to two real-world outbreak events: the 2011 EHEC (enterohaemorrhagic *Escherichia coli*) outbreak in Germany, and the avian influenza outbreak in Cambodia in August 2011.

The aforementioned works show the advantage of using Twitter for detecting real world events focusing on *common and seasonal* diseases, such as, influenza or dengue fever. Existing systems also rely on *particular countries* with a high density of Twitter users, such as, United Sates, United Kingdom or Brazil. However, as seen in Figure 4.6, the tweet volume for EHEC (a non-seasonal disease) is quite pronounced during that outbreak period.

To the best of my knowledge, none of these previous work have focused on an temporal analysis of Twitter data for *general diseases* that are not only seasonal, but also sporadic and that occur in low tweet-density areas like Kenya or Bangladesh, as it will be shown in this work.

An overview of the framework instance is depicted in Figure 4.7. The system aims at generate early warning of outbreaks from twitter messages and to support the temporal

analysis of the twitter extracted events with the events extracted from official reports (ProMED-mail and WHO). The tool provides the ability to visualize and correlate the time series of outbreaks extracted from Twitter with the same event described in the external sources, and compare the results with different granularity of time (days, weeks, months) and space (country, continent, latitude, worldwide).

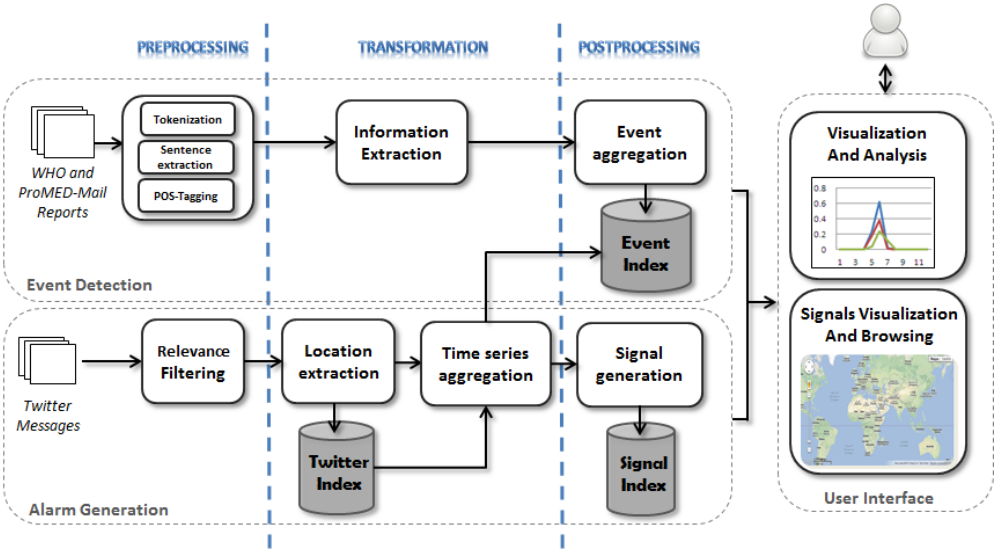

FIGURE 4.7: Framework instance for event extraction and analysis

The system consists of two main blocks: Event Detection and Alarm Generation modules, described in the Sections 4.4.2 and 4.4.3. I recall that in the following sections it will be adopted the terminology about data models described in Section 3.2.1.

4.4.1 Event Model

An event is corresponding to a real-world outbreak, and it can be defined as a quadruple: $e = (v, m, l, t_e)$ described by four attributes that provide information on *who* (victim v) was infected by *what* (disease or medical condition m), *where* (location l) and *when* (time t_e). These four features of an event are extracted from a set of annotated documents. A key aspect of the event model is the extraction of temporal expressions. There are two temporal aspects associated with a disease outbreak e: 1) t_p or the publication time of a document d reporting about e, and 2) t_e or the time of the outbreak, which is the time period that the outbreak has actually token place. In this case, t_e can be determined by time mentioned in d. Note that both t_p and t_e will be used later for the analysis. A set of events E will be extracted automatically from annotated documents, as explained in the Section 4.4.2.

Thus, this event model, if an ongoing outbreak started **last week** and it is first reported in the news of **today**, then the time of the outbreak event I model is last week its, *temporal mentions*, and not today, its *publication time*.

Further, this event model, temporal mentions can be explicit, implicit or relative. Examples of explicit temporal expressions are "May 25, 2012" or "June 17, 2011" that can be mapped directly to dates months, or years on the Gregorian calendar. An implicit temporal expression is an imprecise time point or interval, e.g., "Independence Day 2011" that can be mapped to "July 04, 2011". Examples of relative temporal expressions are "yesterday", "last week" or "one month ago".

4.4.2 Event Detection

The Event Detection module is aimed at extract and aggregate outbreak events. This goal is obtained by procedures executed in a pipeline fashion. The stages of the pipeline consist of: 1) text *Preprocessing* which consist in tokenization, sentence extraction, Part-Of-Speech (POS) tagging; 2) Transformation which consist in named entity recognition and temporal expression extraction; *Postprocessing* with the event aggregation module aimed at aggregate and associate temporal entity to a single event. The entities and spatio-temporal information were extracted using a series of language processing tools, including OpenNLP [116] (for tokenization, sentence splitting and part-of-speech tagging), OpenCalais [112] (for entity recognition) and HeidelTime [139] (for time tagginng). The key aspect of this module is to extract identify and extract events from unstructured raw documents. The approach to event extraction is based on a simple assumption of an event, that is, it is described as a sentence containing a *medical condition* and *geographic expressions*. In this way, there is no need for complex statistical NLP-based techniques or any ontological knowledge for extracting events from a document. A *victim* and the *time of the outbreak* can be identified in the same sentence, or sentences nearby, i.e., sentence context.

Information Extraction. A set of events $\mathcal{E}$ will be extracted from an annotated document $\hat{d} \in \hat{C}$. First, for each annotated document $\hat{d} = (\hat{d}_{ne}, \hat{d}_t, \hat{d}_s)$, I define an event candidate as a sentence $s_i \in \hat{d}_s$ containing **both** a medical condition entity m and a geographic expression l. In other words, I consider a pair of m and l to form an event if they occur in the same sentence s_i, which is more precise than forming events from the cross product of all pairs of medical conditions and locations in a documents which can result in high false positives [140]. Note that, there can be more than one geographic expressions mentioned in s_i, for example:

> *The Health Protection Agency said a 2nd case of anthrax had been confirmed in an injecting heroin user in* **London**, *adding to 2 previous cases in* **England**, *24 in* **Scotland**, *and one in* **Germany**.

The geographic information considered were those belonging to the country-level. Thus, geographic expressions with finer granularity levels like addresses, cities, provinces and states were normalized or mapped to a coarser granularity, such as, a country level using a geo-tagger tool. For example, the geographic expression "London" will be resolved into the country name "England". Finally, I will assume that the identified outbreak is associated to all *distinct* countries recognized in s. Give the above example sentence, it can be identified an anthrax outbreak in 3 different countries: England, Scotland and Germany. The results of this step is a set of *event candidates*: $E_C = \{s_1, \ldots, s_k\}$, where each $s_i \in E_C$ is a sentence associated to a tuple (m, l) of medical condition m and location l.

The next step was to find *victims* and the *event dates* of each outbreak event candidate. For each $s_i \in E_C$, I identify victims and relevant event dates in s_i itself, or in its surrounding context sentences of s_i, i.e., the sentences s_{i-1} and s_{i+1}. The idea of using context sentences is to increase recall of outbreak events being discovered. In order to determine corresponding victims, I assign each recognized victim to the nearest geographic expressions in term of the distance in a sentence. Consider the "anthrax" example above, the numbers of victims associated to **England**, **Scotland** and **Germany** are "2 previous cases", "24" and "one" respectively. Note that the number of victims will be quantified in to a real number, so-called an *case number*, using rules and/or regular expressions. An event date can be identified in the following context sentence:

> *British health authorities repeated a warning to drug users on* **1 March 2010** *that a batch of heroin contaminated with anthrax was probably circulating in Europe, posing a potentially serious health threat.*

The time of an event is defined as *a time period* instead of a time point, because it is an ongoing action. Hence, the starting date and the ending date of an outbreak can be viewed as the earliest date identified, and the publication date of the annotate document $\hat{d}$ respectively, where $s_i \in \hat{d}_s$. In this example, the time of an event t_e is $[tb, tp]$, where tb is the earliest relevant date of s_i, i.e., **1 March 2010**, and tp is the current date wrt. the being considered document d, or the publication date $PubTime(d)$.

If **no** victims and event time can be identified for an event candidate s_i, the s_i is discarded from the set of event candidate. Given the example above, the results from event extraction is a set of "anthrax" outbreaks in 3 different countries. To this end, the final results from this step are event candidates $E = \{e_1, \ldots, e_q\}$, where each $e_i \in E$ is an event represented by a quadruple: $e = (v, m, l, t_e)$ described by four attributes

Bangladesh			Cambodia		
Disease	Cases	Event time	Disease	Cases	Event time
anthrax	6	27 May-02 Jun	Ebola	1	13 May-14 May
anthrax	2	29 May-03 Jun	Ebola	1	13 May-19 May
anthrax	3	03 Jun-07 Jun	Ebola	1	16 May-26 May

TABLE 4.1: Event profiles of *anthrax* in Bangladesh and *Ebola* in Uganda in 2011.

that provide information on *who* v (victim) was infected by *what* (disease or medical condition m), *where* (location l) and *when* (time t_e).

Event aggregation. The set of candidate events E from the previous step were aggregated into a set of outbreak events wrt. a given medical condition m and a particular country l. A *disease event profile* is defined as a set of event candidates associated to a given medical condition m in a particular country l, denoted $dep(l)$.

$$dep(m_i, l_i) = \{e | e = (v, m, l, t_e) \wedge m = m_i \wedge l = l_i\}$$

Examples of the disease event profiles of two countries are shown in 4.1. As seen in the examples, event candidates can be overlapped in time because they might be extracted from a documents writing about follow-up cases, and so on.

Given a disease event profile $dep(m, l)$, the time series of outbreak events $e_{ts}(m, l)$ will be created in the following manner. For each event candidate $e_i \in dep(m, l)$, is performed an iteration over all dates in the event time period $t_{e,i}$ and assigned the case number v_i to each date $t \in t_{e,i}$. If there is any overlap date t_k between two event candidates e_i and e_j, the case volume of t_k will be the sum of the case numbers v_i and v_j of event candidates e_i and e_j respectively. It is important to note that a case volume is roughly *estimated* and it is heavily depend on the accuracy of annotation tools.

The result of this process is a *set of outbreak events* that occurred in different time and place for a set of diseases, where each outbreak will be associated with the number of victim/suspected cases. The extracted events were manually verified with the input of domain experts in order to filter out irrelevant reports, such as, those containing updates or discussion about the characteristics of a disease. The extracted events were indexed and used as *ground truth* for further analysis. Since Twitter data is highly ambiguous and noisy, it was applied a filtering technique to filter out *non-relevant* tweets (as Described in Section 4.4.3). Then, relevant tweets will be indexed also for further analysis.

4.4.3 Alarm Generation

The Alarm Generation module aims to identify relevant tweets (those that are matched with an outbreak event) in order to be able to generate an alarm when an anomaly is

Disease	MedISys	Urban Dictionary
mumps	multi+frontal+massively, programming+system	something is really cool or exciting
dengue	concert, film, novel, book, music, band, movie, album	something is disgusting ghetto, nasty, or sketchy
scarlet fever	N/A	the condition of being a redophile
yellow fever	independent, film, festival	preference for Asian women

TABLE 4.2: Examples of negative keywords/phrases for a disease name collected from MedISys and Urban Dictionary.

detected within the timeseries of twitter messages associated with the event. Twitter data is highly ambiguous and noisy. A disease name mentioned in a tweet can have many contexts that are not relevant to an outbreak. For example, irrelevant tweets for epidemic analysis are: (a) *"A two hour train journey, Love In the Time of Cholera."* or (b) *"I liked a @YouTube video http://youtu.be/... a Metallica, Megadeth, & Anthrax - Helpless"*. Both tweets mention an infectious diseases, namely: Cholera and Anthrax, but their context is literature and music, respectively. Additionally, tweets about vaccine, marketing campaigns, or ironic/jokes are considered non-relevant.

The Relevance Filtering block is aimed to distinguish relevant tweets from non-relevant ones by using keyword feature, as described below. Referring to the tweet model described in Section 3.2.2 a tweet, tw, is considered non-relevant if it contains one of the negative keywords wrt. a disease name or a medical condition m, where m is a term in the contents of tweet tw_{text}.

Positive keywords associated to diseases are pathogen (e.g., Streptococcus pyogenes) and symptoms (e.g., sore throat, fever, bright red tongue with a strawberry appearance, rash, bumps, itchy, and red streaks). The negative keywords associated to diseases are collecterd from two freely-available resources: 1) MedISys [3] providing a list of negative keywords created by medical experts, and 2) Urban Dictionary [4] a Web-based dictionary of slang, ethnic culture words or phrases. Examples of negative keywords/phrases for disease names are shown in Table 4.2. For each tweet tw, it is considered non-relevant if it contains one of the negative keywords wrt. a disease name or a medical condition m, where m is a term in the contents of tweet tw_{text}.

The location associated with a tweet is made through the use of three information listed in order of importance: 1) mention of the location (city, country, etc.). Contained in the text 2) if present, the geolocation information (latitude and longitude) message, and 3) location indicated in the user profile of the tweet. All places recognized are defined through the use of API Yahoo! PlaceFinder. The relevant tweets with location information are indexed for further analysis. The module of TimeSeries Aggregation is responsible for defining time series of events indexed while the modulus of Signal

[3] http://medusa.jrc.it/medisys/homeedition/en/home.html
[4] http://www.urbandictionary.com/

Generation takes care to identify anomalies in the time series and to generate alarm signals. The generation of signals arising from tweet is not an easy task because of two main challenges: (i) the messages are highly ambiguous, because of the few characters of which are compounds tweets the steps of filtering of messages relevant are partly compromised because these errors propagate in the results, (ii) the characteristics of infectious diseases are very dynamic in time and space and thus their behavior varies widely between the different regions and periods of the year.

Chapter 5

Evaluation

In Chapter 4 the *Document Processing Framework* has been instantiated and as a result an architecture for heterogeneous and multi-language data management was defined. This architecture is aimed to support the medical domain actors to accessing and retrieving useful information. In this chapter it will be provided experimental results that proof the answers to the research question provided in the Introduction (Section 1.1:

Question 1: Is it possible to automate processes for data classification of paper documents? And, how?

Question 2: Can be the access policies applied automatically for fine-grain protection of information?

Question 3: How external sources can be exploited for complementing traditional information?

5.1 Medical Record Classification with Clustering Ensemble

For the experimental campaign it was adopted a corpus composed by real, paper, medical records belonging to four different hospital departments that was digitalized by means of a state-of-the-art Optical Character Recognition (OCR) technique. As described in section 4.2 it were built three different classes of vector space models considering respectively *terms*, *lemmas* and *concepts*, that are the doc-matrices (A). These vector space models were used to generate, for each one of them, $L = 12$ different instances of the X-means clustering algorithm. The dataset used to validate the adopted strategy is

made up of 143 documents, which are scans of medical records, obtained from different Italian hospitals.

It is worth to note that since the dataset used is composed by digitalized medical records, they are affected by noise. The noisy terms are discarded by means of the preprocessing phase of the semantic methodology and so the documents are represented by the terms correctly recognized by the OCR procedure. It implies that the document representation in the vector space considers a subset of the original terms that occurs in the medical records.

These records were previously organized on the basis of department membership as follows: *Cardiology*: 41 documents; *Intensive Case*: 40 documents; *General Surgery*: 40 documents; *Oncology*: 22 documents. On the three doc-feature matrix A it was evaluated the proposed approach, obtaining the results reported in the Table 5.1.

Criteria	Rand index	NMI index	Number of Clusters
Indexes mean and standard deviation among the 12 partitions			
I Criterium (Terms)	0.5523 ± 0.0035	0.1946 ± 0.0055	2.8333 ± 0.8788
II Criterium (Lemmas)	0.5042 ± 0.0058	0.2507 ± 0.0039	2.5833 ± 0.4470
III Criterium (Concepts)	0.4925 ± 0.0044	0.2344 ± 0.0026	2.5000 ± 0.4545
Criteria Combination			
Terms + Lemmas	0.7164	0.3234	5
Lemmas + Concepts	0.7313	0.3674	7
Terms + Concepts	**0.7324**	**0.3787**	7
Terms + Lemmas + Concepts	0.7286	0.3491	7

TABLE 5.1: Model evaluation through the Rand index, Normal Mutual Information (NMI) index and the number of clusters; the best cases are highlighted in bold

Although the use of the III Criterion allows us to make documents' partition by topic, it introduces noise, making the partitions generated worse than the ones obtained by using only the I or II Criterion. On the other hand, the usage of this information combined with the I or the II Criterion performs better. In particular, the best results are obtained by combining features coming from the I Criterion and the III Criterion.

5.1.1 Discussion

Research Question 1: *Is it possible to automate processes for data classification of paper documents? And, how?*

In previous section I proposed a methodology for automatic document categorization based on the adoption of clustering ensemble technique. I proposed a feature selection based on semantic processing for represent data in the vector space model. From this were built tree different classes of vector space models considering respectively terms,

lemmas and concepts. Those vector space models were adopted document clustering. I combined different results of X-means clustering algorithm executed with three different vectors space models which include syntactic and semantic content representation.

For the experimental campaign it was adopted a corpus composed by real medical records that were digitalized by means of OCR. Then I used the scanned data to perform the experiments. Since the dataset used is composed by digitalized medical records, the data used for testing present some noise. The noisy terms are discarded by means of the preprocessing phase of the semantic methodology and so the documents are represented by the terms correctly recognized. It implies that the document representation in the vector space considers a subset of the original terms that occurs in the paper medical records. In the experiments I adopted the dataset in order to build the vector space model classes. For each classes, I built two kind of vector space model: in the first case was considered the dataset composed by the whole medical record set; in the second case were considered only the first two pages of each medical record that report a summary of the whole medical record.

The results showed that although the use of concepts allows to make documents' partition by topic, it introduces noise, making the generated partitions worse than the ones obtained by using only lemmas or terms. On the other hand, the usage of semantic information combined with the syntactical ones allowed to improve the obtained results.

5.2 Fine grain access policy for document protection

In order to process documents belonging to the E-Heath domain, the document processing framework has been instantiated by means of input parameters o_i selection and corresponding techniques. In this way specific tools have been selected in order to implement each module. The system accepts in input the corpus (made of unstructured medical records) and performs the formalization activities in order to structure it. In this way the resources to protect, representing the objects of the security rules, are easily located.

In order to illustrate the processing phases, let's consider a fragment of an Italian medical record:

> "La Signora si presenta con un anamnesi di precedenti ricoveri presso differenti reparti di questo ospedale. Inquieta ed a tratti aggressiva, manifesta un forte stato d'ansia e dolori allo stomaco. Vista la storia clinica di patologie ansiogene del paziente, le sono stati somministrati 10mg di Maalox."

Although the example is formulated in Italian, the concepts to whom the relevant terms refer to will be indicated in English.

The fragment states that "the patient presents a history of previous admissions in different departments of a hospital. Restless and aggressive, shows a strong state of anxiety and stomach pain. Given the patient's anxiety-inducing conditions, she was given 10mg of Maalox".

Once the terms of this fragment were extracted by means of *Preprocessing* module (as described in Section 3.3), the *Transformation* module extracts the relevant terms using, as described above, statistical measures; all the terms having a *tf-idf* value over an established threshold are selected: "paziente"(4.1), "ansia"(4.2), "dolori allo stomaco"(3.8), "aggressiva" (3.1), "storia clinica"(4.8), and "Maalox" (4.7). These terms are then linked to the synsets to which they belong. Each synset refers to a concept, and each concept is then associated to a document section as summarized in Figure 5.1.

Extracted Terms	**Associated Synset**	**Concept**	**Associated Section**
paziente	ammalato, degente, malato, paziente	Patient	Invest. and Diagnosis
ansia	ansioso, ansiogeno, anxiety	Anxiety	Invest. and Diagnosis
dolori allo stomaco	dolori_allo_stomaco, mal_di_somaco,	Stomach Pain	Invest.and Diagnosis
aggressiva	aggressivo, aggressiva, aggressive	Aggressive	Patient Status
storia clinica	storia clinica, patient history	Patient History	Patient Status
Maalox	Maalox	Maalox	Therapy

FIGURE 5.1: Association between extracted terms and corresponding concepts

In the example I obtained the following concepts associated to the extracted terms: "Patient", "Anxiety" and "Stomach Pain", "Aggressive", "Patient History" and "Maalox".

The presence of concepts "Patient", "Anxiety" and "Stomach_Pain" in the underlined part of the example, and the absence of concepts belonging to the other sections, constitutes the features by which the subsection of the fragment under analysis will be classified as "Investigation and Diagnosis" section. The relevant concepts are structured in RDF format and the list of identified resource are coded in the HL7 standard for medical records by the *Postprocessing* module (see an example in Appendix A).

The security policy is made of a set of rules structured as follows.
A rule is as a triple $\langle s_j, a_i, r_k \rangle$ where $s_j \in S$, $a_i \in A$, $r_k \in R$ and:

- $S = \{s_1, ..., s_m\}$ is the set of the actors s_j that can access to the medical record,
- $R = \{r_1, ..., r_n\}$ is the set of all resources (sections) r_i belonging to the medical record,

- $A = \{a_1, ..., a_h\}$ is the set of actions that can be performed by an actor $s_j \in S$ on a resource $r_i \in R$.

For each resource, a subset of rules belonging to the applicable policy set is available.

So, given a resource $r^* \in R$, all the possible rules, denoted as L_{r^*}, belonging to the Policy will be retrieved; by definition:

- $L_{r^*} = \{\langle s_j, a_i, r^* \rangle \, | r^* \in R, s_j \in S^* \subseteq S, \;\; a_i \in A^* \subseteq A\}$ is the set of all allowed combinations of (subjects,actions) on the resource r^*.

This kind of resource can be accessible only by those people having proper rights. In a role-based access control mechanism, these rights are associated to a role and they are usually assigned by security administrators according to governmental laws (on privacy in e-health, for example) and enterprise regulations. I designed security policies for this domain, located roles were: *Doctors, Administrative Managers, Nurses* and *Patient*. As a sample case, for each role, the considered actions are: "read" and "write". In the illustrated example, being the selected text fragment identified as belonging to the "Investigation and Diagnosis" section, the security mechanism must retrieve applicable rules and enforce them; a set of possible rules are:

R1: $\{Doctor, \, Investigation\,and\,Diagnosis, \, (Read, \, Write, \,)\}$
R2: $\{Nurse, \, Investigation\,and\,Diagnosis, \, (Read, \, \neg Write)\}$
R3: $\{Administrative\,Manager, \, Investigation\,and\,Diagnosis, \, (\neg Read, \, \neg Write)\}$

For example, R1 states that the Doctor can Read and Write on resources of type "Investigation and Diagnosis" while a Nurse can just Read it (R2) and an Administrative Manager cannot access it (R3). In Figure 5.2 the postprocessing module is illustrated.

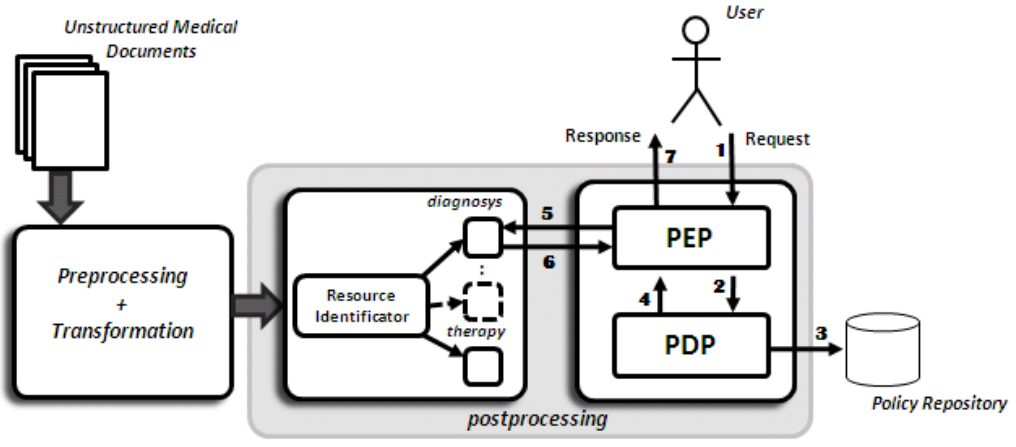

FIGURE 5.2: Postprocessing module for resource access control

After resource identification and coding, the policy enforcement is implemented through the standard eXtensible Access Control Markup Language (XACML) architecture [147]; it includes the Policy Enforcement Point (PEP) that captures the request from a generic user with an associated role and the Policy Decision Point (PDP) that evaluates the request on the basis of the policy included in the policy repository.

5.2.1 Discussion

Up to date, many systems are based on document management applications and cannot benefit of new design techniques to structure data because of the presence of old unstructured documents written by doctors, lawyers, administrative people and so on. Indeed, documents, especially the old ones, are just digitalized and made available to users. Among the other limitations, this prevents access control mechanisms from enforcing fine-grain security policies. In Section 5.2 I proposed an innovative framework that is based on a semantic methodology [7] to transform unstructured data in a structured way by extracting relevant information and to identify critical resources to protect. It was also illustrated its adoption on a simple case study to put in evidence the potentiality of the proposed approach from a security perspective by formalizing e-Health record and defining fine grained access control rules.

5.3 Event Extraction from Twitter

In this section, I'll present evaluations on outbreak event extraction (Section 5.3.1) and on signals triggered from Twitter data (Section 5.3.2).

5.3.1 Outbreak Event Analysis

As described in Section 4.4, the framework instance for event extraction allows a user to visualize and analyze the temporal development of an outbreak event in Twitter by comparing with the outbreak information automatically extracted from official sources. The data are represented as an interactive, zoomable plots of time series that can be easily explored by the user. The charts are implemented by employing the Dygraphs JavaScript visualization library[1]. Given a time series graph, it is possible to explore and display values on mouse over allowing the user to analyze time series data in a particular time period. In addition, it is also possible to evaluat the cross-correlation coefficient (CCF), or a statistical method to estimate how variables are related at different time

[1] http://dygraphs.com/

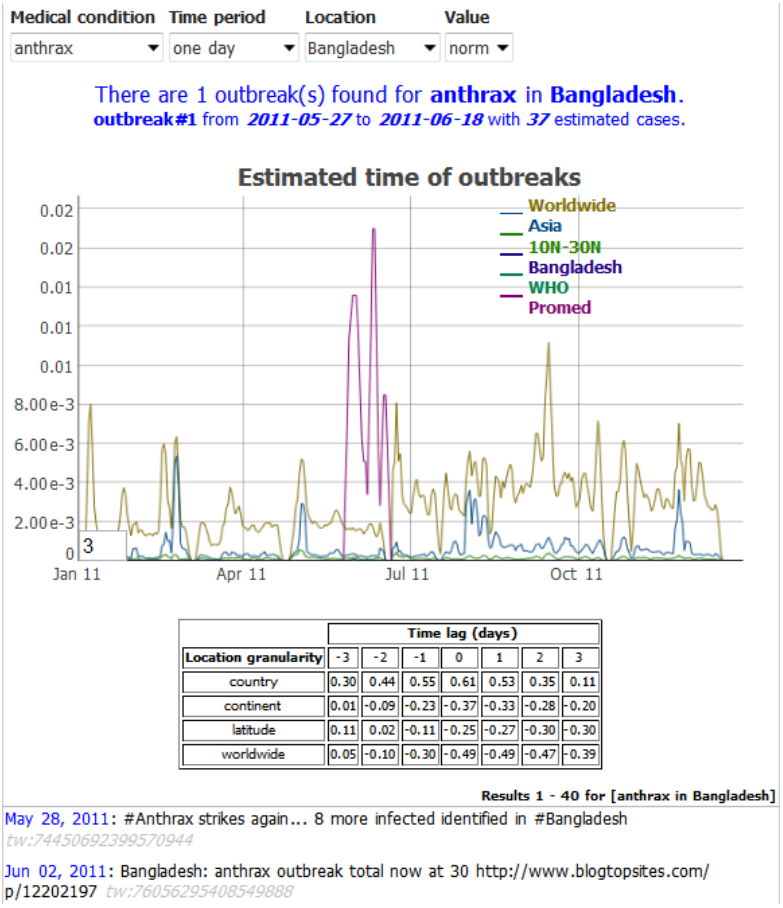

FIGURE 5.3: Temporal development of the 2011 anthrax

lags. This measurement can be interpreted as the similarity between two time series in volume, with consideration of time shifts. That is, the CCF value indicates whether there is a correlating trend in Twitter wrt. a real-world outbreak event. Figure 5.3 demonstrates the temporal analysis of the anthrax outbreak occurred in Bangladesh in 2011. The interface allows the user to filter the time series visualization considering different granularities of *time* and *location*. Moreover, a moving average parameter (in days) can be adjusted by the user. In the Figure 5.3, the Twitter time series corresponding to all location granularities are shown, given a one-day time granularity and a smoothing parameter of 3 days. Below of the graph, the cross correlation results of 3-day time lag are also displayed.

5.3.1.1 Cross Correlation Coefficient

In time series analysis, the cross-correlation coefficient (CCF) is a statistical method to estimate how variables are related at different time lags. That is, the CCF value at time t between two time series X and Y indicates the correlation of the first series with respect to the second series shifted by a time amount t, e.g., in days or weeks. A common

measure for the correlation is the Pearson product-moment correlation coefficient. The CCF between two time series describes the normalized cross covariance and can be computed as:

$$ccf(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[(X - E\left[X\right])(Y - E\left[Y\right])\right]}{\sigma_x \sigma_y}$$
$$= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

where x_i and y_i are values at time t_i of X and Y, $\bar{x}$ and $\bar{y}$ are the means values, and σ_X and σ_Y are the standard deviations. In our case, two time series X and Y are corresponding to the time series of Twitter and a real-world outbreak event respectively. The $ccf(X,Y)$ function has values between -1 and +1, where the value ranges from 1 for perfectly correlated results, through 0 when there is no correlation, to -1 when the results are perfectly correlated negatively. This measurement cab be interpreted as the similarity between two time series in volume, with consideration of time shifts. That is, the CCF value indicates whether there is a *correlating trend* in Twitter wrt. a real-world outbreak event.

5.3.1.2 Matching Tweets

To perform an analysis, it is needed to *match* tweets with an outbreak event using keywords: medical condition and location. That is, a tweet will be matched with an outbreak if it contains a medical condition and its location is the same as that of the outbreak event. A location associated to each tweet will be normalized into two different granularities with respect to a geographic concept hierarchy: country and continent. The intuition of using different geographic granularities is that the public attentions depend on their geographic distance from an outbreak event. For example, people might talk or share their opinions about an ongoing outbreak in a neighborhood country because they are concerned that the outbreak can spread into their country. Consequently, I consider not only a *country-level* location, but also a *continent-level* location. Thus, given an outbreak event e, a tweet tw is matched with e if $country(tw_{loc}) = l_e$ or $continent(tw_{loc}) = l_e$, where $country(l)$ is a mapping function from a location l into a country-level location, and $continent(l)$ is a mapping function from a location l into a continent-level location.

It were analyzed Twitter data collected during the period of 1st January 2011 to 31 December 2011. In this period were collected 112.134.136 tweets containing one of the

1000 English keywords related to a list of medical conditions and symptoms provided by the Robert Koch Institute (RKI)[2].

From this set of twitter post were extracted those related to all the medical condition reported in table 5.4 obtaining 1.468.504 tweets from 744.822 distinct users. In Table 5.2 are shown the percentage of locations related to tweets of the dataset used.

Twitter	
Location Entity	28%
Location Author	78%
Geo Location	1%

TABLE 5.2: Twitter collection

5.3.1.3 Identifying Relevant Time

The task of identifying relevant time can be regarded as a *classification problem*. That is, it will be determined whether a temporal expression is **relevant** or **irrelevant** for a given event. It was employed a machine learning method for learning the relevance of temporal expressions using three classes of features: sentence-based, document-based and corpus-specific features. Note that, the proposed features can be applied for the similar task in a generic domain as well.

Sentence-based Features. Given a temporal expression t_e, the values of features are determined from the sentence s_i containing t_e, where s_i is extracted from an annotated document $\hat{d}$. For this class, 13 features are proposed, namely, *senLen*, *senPos*, *isContext*, *cntEntityInS*, *cntTExpInS*, *cntTPointInS*, *cntTPeriodInS*, *entityPos*, *entityPosDist*, *TExpPos*, *TExpPosDist*, *timeDist*, and *entityTExpPosDist*. The intuition is to determine the relevance of temporal expressions by considering *the degree of relevance of their corresponding sentences* with respect to a given event. For example, a sentence that is too long or too short is likely to be less relevant, and a sentence containing too many of geographic expressions is possibly irrelevant or less specific to an event. The granularity type of time mentioned in a sentence can also indicate the relevance to a particular event, e.g., a time point should be more relevant than a time period because it is more precise/accurate.

The first feature *senLen* is a score of the length (in characters) of s_i normalized by the maximum sentence length in $\hat{d}$. The feature *senPos* gives a score of the position of s_i in $\hat{d}$ normalized by the total number of sentences in $\hat{d}$. The feature *isContext* indicates whether s_i is a context sentence or not. The feature *cntEntityInS* is a score

[2]The Robert Koch Institute is the German central federal institution responsible for disease control and prevention

of the number of occurrences of entities in s_i normalized by the maximum number of entities in any sentence $s_i \in \hat{d}$. The feature *cntTExpInS* is a score of the number of temporal expressions in s_i normalized by the maximum number of temporal expressions in any sentence $s_i \in \hat{d}$. The features *cntTPointInS* and *cntTPeriodInS* are scores of the numbers of time points and time periods in s_i normalized by the number of temporal expressions in s_i respectively. In other words, the two features take into account time granularities, such as, either a point or a period of time.

The feature *entityPos* is an average of scores of the positions (in character) of entities in s_i normalized by the length of s_i. The feature *entityPosDist* is an average of scores of the position distance between all pairs of entities in s_i normalized by the length of s_i. The feature *TExpPos* is an average of scores of the positions (in character) of temporal expressions in s_i normalized by the length of s_i. The feature *TExpPosDist* is an average of scores of the position distance between all pairs of temporal expressions in s_i normalized by the length of s_i. The feature *timeDist* is an average of scores of the distance *in time* for all pairs of temporal expression in s_i. The assumption is that the further distance two time expressions have, the less they are related. The final feature is *entityTExpPosDist*, which is an average of scores of the position distance between all pairs of (entity, time) in s_i normalized by the length of s_i. Note that, this feature is only applicable when s_i is a *not* context, but the original sentence of an event. The value of all features is normalized to range from 0 to 1.

Document-based Features. Belong to this category five features that are determined at the document level, namely: *cntEntityInD*, *cntEntitySen*, *cntTExpInD*, *cntTPointInD*, *cntTPeriodInD*. In general, the proposed features are aimed at capturing the **ambiguity** of a document mentioning about a given event. These features can be computed off-line because they are independent from an event of interest.

The first feature *cntEntityInD* is a score of the number of occurrences of entities in $\hat{d}$ normalized by the total number of sentences in $\hat{d}$. The feature *cntEntitySen* is a score of the number of sentences containing at least one entity normalized by the total number of sentences in $\hat{d}$. The feature *cntTExpInD* is a score of the number of temporal expressions in $\hat{d}$ normalized by the total number of sentences in $\hat{d}$. The feature *cntTPointInD* is a score of the number of time points in $\hat{d}$ normalized by the total number of temporal expressions in $\hat{d}$. The feature *cntTPeriodInD* is a score of the number of time periods in $\hat{d}$ normalized by the total number of temporal expressions in $\hat{d}$. Similar to the previous class, the values of all features is normalized to range from 0 to 1.

Corpus-specific Features. This class of features is based on heuristics with respect to a particular document collection. Temporal expressions are considered **non-relevant** if

they are mentioned in question or negative sentences as well as those related to commercial or vaccinate campaigns. It was manually build *negative keywords* set corresponding to different aspects mentioned above. The feature *isNeg* is 1 if a sentence s_i contains any term in negative keywords, and it is 0 otherwise. In addition, temporal expressions are **not** related to a given event if they refer to a history of an outbreak, e.g., some statistics in the past. Thus, the feature *isHistory* is 1 if a sentence s_i contains any term related to historical data (e.g., "statistic", "annually", "past year") and it is 0 otherwise.

5.3.1.4 Evaluation of Relevant Time

The document collection considered in this work consists of official medical reports posted all over the year 2011 and provided by two different authorities: the World Health Organization (WHO) [151] and ProMED-mail [121]. The reports contain information about outbreaks and public health treats, which were moderated by medical professionals worldwide. The number of documents and sentences collected for ProMED-mail are 2,977 documents and 95,465 sentences; whereas for WHO only 59 documents were reported resulting in 761 sentences. The text annotation required a series of language processing tools, including OpenNLP [116] (for tokenization, sentence splitting and part-of-speech tagging), OpenCalais [112] (for named entity recognition) and HeidelTime [139] (for temporal expression extraction).

The dataset was created by manually selecting 25 infectious diseases (medical conditions) by medical professionals, and outbreak events were extracted with respect to the selected diseases. The main goal is to evaluate the proposed method for identifying the relevant temporal expressions of a given event. Thus, it was asked human assessors to evaluate event/time pairs (e.g., relevant or non-relevant) using 3 levels of relevance: 1 for *relevant* to an event, 0 for *irrelevant* to an event or *incorrect* tagged time, and -1 for *unknown*. The *incorrect* tagged time is an error produced by the annotation tools. More precisely, an assessor was asked to give a relevance score $Grade(e, t_e)$ where $Grade(e, t_e)$ is a pair of an event e, and a temporal expression t_e. When t_e is a time period, i.e., containing two dates, an assessor has to give judges to both dates. Hence, an event/time pair (e, t_e) is relevant if and only if there is at least one relevant date, and it is **non-relevant** if all dates are non-relevant. Finally, assessors evaluated about 3500 event/time pairs.

The Weka implementation [154] was used for modeling the relevant time identification as a classification task, which was learned using several algorithms: decision tree (J48), Naïve Bayes (NB), neural network (NN) and SVM, using 10-fold cross-validation with 10 repetitions. I measured statistical significance using a t-test with $p < 0.05$. In the table, bold face indicates statistically significant difference from the respective baseline.

Feature	*J48*	*NB*	*NN*	*SVM*
senLen	**.65**	.59	.58	.58
senPos	**.62**	.58	.58	.58
isContext	.58	.58	.58	.58
cntEntityInS	.59	.53	.58	.57
cntTExpInS	.61	.55	.58	.57
cntTPointInS	.60	.59	.59	.58
cntTPeriodInS	.60	.59	.59	.58
entityPos	.65	.58	.58	.57
entityPosDist	**.65**	.58	.58	.57
TExpPos	.58	.58	.58	.58
TExpPosDist	**.59**	.59	.59	.58
timeDist	.58	.58	.49	.58
entityTExpPosDist	.57	.58	.58	.58
cntEntityInD	.62	.58	.58	.57
cntEntitySen	.62	.58	.58	.57
cntTExpInD	.63	.52	.58	.57
cntTPointInD	.61	.58	.58	.58
cntTPeriodInD	.61	.58	.58	.58
isNeg	.58	.58	.58	.58
isHistory	.58	.58	.58	.58
senBased	.66	.55	.59	.59
docBased	.68	.58	.61	.60
corpusBased	.58	.58	.58	.58
ALL	.69	.55	.61	.63

TABLE 5.3: Accuracy of relevant time identification.

Classification results. The baseline method for relevant time classification is the majority classifier. The accuracy of the baseline is 0.58. Table 5.3 shows the accuracy of different classification algorithms on each feature. The combination of all features within a particular class is denoted *senBased*, *docBased*, and *corpusBased* respectively. *ALL* is the combination of all features among different classes.

The overall results show that decision tree (J48) is the best among other classification algorithms. In general, many of sentence-based features improved the accuracy of baseline significantly. The features *senLen* and *entityPosDist* perform best with accuracy=0.65. While the features in document-based class obtained high accuracy, but they did not significantly improve the baseline. The worst performing features are those from corpus-specific class. The combination of different features gained high accuracy but did not significantly outperform the baseline.

5.3.2 Generating Signals from Twitter

Generating signals using Twitter data is a difficult task because of two main challenges:

Highly ambiguous and noisy data. Time series data created from tweets and used for signal generation are noisy, incomplete and sparse, in part, from propagation of errors within the previous stages or nature of the data. Noise can be caused by spurious events in which an entity is correctly detected, but its role is not. For example, irrelevant tweets

for epidemic analysis are: 1) *"A two hour train journey, Love In the Time of Cholera."* or 2) *"I liked a @YouTube video http://youtu.be/... a Metallica, Megadeth, & Anthrax - Helpless"*. Both tweets mention infectious diseases, namely: Cholera and Anthrax, but their context is literature and music, respectively.

Incomplete or sparse time series data implies that instances of an event are missing or under-reported. This may occur due to: 1) the presence of processing errors - an acronyms or abbreviations not recognized as medical conditions; 2) the fact that people who are actually suffering do not tweet; 3) the tweets which contain these mentions have not been collected by the system, i.e., based on the imbalance between the type of tweets collected (e.g., personal versus news tweets); and 4) the minimum required entity types are not present. Sparse time series data refers specifically to low aggregation counts, which impact the anomaly detection algorithm [5].

Temporal and spatial dynamics of diseases. The characteristics of infectious diseases are highly dynamic in time and space, and their behavior varies greatly among different regions and the time periods of the year. Some infectious diseases can be rare or aperiodic, while others occur more periodically. In addition, various diseases have different transmission rates and levels of prevalence within a region. For example, cholera infections vary greatly in frequency, severity, and duration. On the one hand, in some regions historically, only sporadic outbreaks occur in areas, such as, parts of South America and Africa. On the other hand, even in areas where cholera infections are endemic (the South Asian countries of Bangladesh and India) the epidemic levels change dramatically from one year to the next [64].

Given the imperfect time series data, in this section is analyzed the extent to which it is possible to trust signals that have been generated for early warning. To this end, were followed the guidelines given by Collier [39] and Khan [92].

Studying the usefulness of Twitter data in an early warning task requires real-world outbreak statistics. I build an outbreak ground truth by relying upon ProMED-mail , a global reporting system providing information about outbreaks of infectious diseases. In total were collected 3,056 ProMED-mail reports and identified 14 different outbreaks occurring during year 2011 as ground truth. The outbreaks are detailed in Table 5.4.

An important aspect of this work is that I consider the duration of each outbreak by analyzing temporal expressions in a ProMED-mail document, unlike aforementioned work [39] that assumes the publication date of a document as the estimated relevant time of an outbreak. In particular, I determined the starting date of a disease by looking at the first ProMED-mail post, and the ending date was related to the last ProMED-mail publication, for that disease-location. One reason for doing this is that the events

in ProMED-mail undergo moderation, so there is often a delay between the time of the actual outbreak and the publication date of the related report. However it is worth noting that this strategy gives a good confidence only on the beginning date of the outbreak. Indeed, the absence of further ProMED-mail posts does not necessarily mean an end of the outbreak, but just that there was no significant news in which it was reported.

ID	Disease	Country	Event period
1	anthrax	Bangladesh	[11-May,18-Jun]
2	anthrax	India	[03-Jun,22-Jun]
3	botulism	Finland	[17-Oct,01-Nov]
4	botulism	France	[01-Sep,10-Sep]
5	cholera	Kenya	[11-Nov,03-Dec]
6	ebola	Uganda	[13-May,30-Jun]
7	ehec	Germany	[05-May,30-Jun]
8	leptospirosis	Denmark	[02-Jul,23-Jul]
9	leptospirosis	Philippines	[27-Jun,15-Jul]
10	mumps	Canada	[10-Jun,17-Aug]
11	mumps	United States	[27-Sep,11-Oct]
12	norovirus	France	[16-Jul,25-Jul]
13	rubella	Fiji	[26-Jul,09-Aug]
14	rubella	New Zealand	[15-Aug,19-Aug]

TABLE 5.4: List of 14 outbreaks: each outbreak is represented by ID, disease (or a medical condition), country and the duration of the event.

5.3.2.1 Biosurveillance Algorithms

In order to detect outbreak events for early warning, were exploited different state-of-the-art algorithms as anomaly detectors in disease-related Twitter messages: **C1**, **C2**, **C3**, F-Statistic (**FS**), Experimental Weighted Moving Average (**EWMA**) and Farrington (**FA**) [21, 69]. The objective of biosurveillance is the detection of emerging incidence clusters in time of a health related event. The surveillance algorithms used are well documented in the disease aberration literature, e.g., [21, 82, 92]. These algorithms were applied in order to detect aberration patterns in the time series when the volume of tweets mentioning a medical condition or medical condition-location pair exceeds an expected threshold value. Below I provide details of the biosurveillance algorithms used for early detection.

Early Aberration Reporting System (EARS) algorithms: **C1**, **C2**, and **C3** [82]. These algorithms compute a test statistic on day t as follows:

$$S_t = \max(0, (X_t - (\mu_t + k\,\sigma_t))/\sigma_t) \tag{5.1}$$

where X_t is the count on day t, k is the shift from the mean to be detected, and μ_t and σ_t are the mean and standard deviation of the counts during the baseline period. For C1, the baseline period is $(t-7,\ldots,t-1)$; for C2 the baseline is $(t-9,\ldots,t-3)$. The test statistic for C3 is the sum of $S_t + S_{t-1} + S_{t-2}$ from the C2 algorithm (Eq. 5.1). The constant k determines how sensitive is the algorithm to generate a signal. With a lower value of k, the algorithm becomes more sensitive as it will trigger an alarm with less of a deviation from the mean of the process.

F-statistic [35] is computed as: $S_t = \sigma_t^2 + \sigma_b^2$, where σ_t^2 approximates the variance during the testing window and σ_b^2 approximates the variance during the baseline window. Their calculation is as follows:

$$\sigma_t^2 = \frac{1}{n_t} \sum_{test}^{n_t} (X_t - \mu_b)^2$$

$$\sigma_b^2 = \frac{1}{n_b} \sum_{test}^{n_b} (X_t - \mu_b)^2$$

Exponential Weighted Moving Average (EWMA) model [92] is provided for a non-uniformly weighted baseline by down-weighting counts that are on days further from the target day. The smoothed daily counts were calculated as:

$$Y_1 = X_1; \;\; Y_t = \omega\, X_t + (1-\omega)\, Y_t - 1$$

and the test statistic was calculated as

$$S_t = (Y_t - \mu_t)/[\sigma_t * (\omega/(2-\omega))^{1/2}]$$

where $0 > \omega > 1$ is the smoothing constant, and μ_t and σ_t are the mean and standard deviation for the baseline window, which was set to $(t-15,\ldots,t-5)$.

The **Farrington** detection algorithm. The aim of this algorithm is to predict the observed number of counts based on a subset of the historic data by extracting reference values close to the week under investigation and from previous years. The algorithm fits an over dispersed Poisson generalized linear model (GLM) with log-link to the reference values. The Farrington algorithm is consider to be a robust and fast method applicable for the routine monitoring of weekly reports on infections for many different pathogens. Please refer to [69] for further details.

An alarm is triggered if the test statistic exceeds a threshold value, which is determined experimentally. The larger the amount by which the threshold is exceeded, the greater the severity of the alarm.

The common parameters and the setting adopted for each algorithm are:

- *training window*: number of days before the analyzed outbreak period. The training window is used to compute statistics for anomaly prediction. Window sizes considered have the following values: {4, 5, 8, 9, 10},
- *buffer*: guard days before the target day being assessed; used together with training window. Buffer sizes used in our analysis are 1 and 2 days, and
- *alarm threshold*: a threshold value used to assess if the target day presents an anomaly. In this study I evaluated the algorithms using the following threshold values: {0.1, 0.2, . . . , 0.8, 0.9}.

5.3.2.2 Evaluation Metrics

The metrics used to assess generated signals are *Sensitivity*, *Predictive Positive Value* and *F-measure*. Sensitivity refers to the proportion of true signals correctly detected by a surveillance algorithm. Predictive Positive Value (PPV) is the proportion of signals generated by the algorithm that actually corresponds to an outbreak. A low value of PPV indicates high false positive outbreaks leading to unnecessary intervention, whereas a high PPV value will lead to fewer misdirected sources of information during the investigation of a potential outbreak. In order to measure the system performance in a single metric, I use F-measure that is the harmonic mean of sensitivity and PPV. All metrics can be computed as follows:

$$Sensitivity = TP/(TP + FN)$$

$$PPV = TP/(TP + FP)$$

$$F\text{-}measure = 2 * (Sensitivity * PPV)/(Sensitivity + PPV)$$

where TP (true positive) is the number of true signals generated in the presence of an actual outbreak; FN (false negative) corresponds to the number of true signals not generated, and FP (false positive) is the number of signals generated in the absence of an outbreak.

Comparison Method. In this part is presented the study of which algorithm and parameter settings perform better in terms of F-measure by varying an evaluation window

including the outbreak period and a period outside that contains as many days with at least one disease-related tweet as in the outbreak period. I consider an alarm triggered outside the outbreak period as a false alarm. I look for the best configuration of parameters that maximized the F-measure performance using grid search.

5.3.2.3 Evaluation of Biosurvelliance Algorithms

In this section, I first discuss some observation in categorizing outbreak time series data. The results of the analysis can provide useful insights about properties of the events. I identify that the real-world outbreaks reflected in Twitter can be characterized according to two dimensions: 1) oscillation seen as the frequency at which the curve spikes; and 2) volume (or magnitude) of daily count of twitter messages, sinks or slopes. Representative samples of outbreak-related Twitter time series data are depicted in Figures 5.4, 5.5, 5.6 and 5.7.

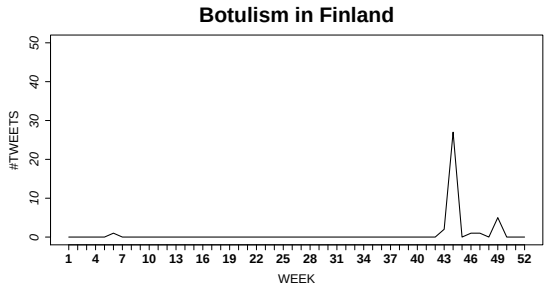

FIGURE 5.4: Low Oscillation and Low Magnitude

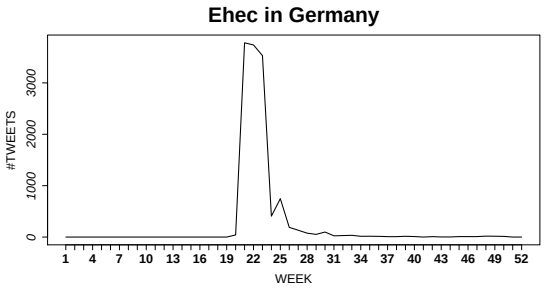

FIGURE 5.5: Low Oscillation and High Magnitude

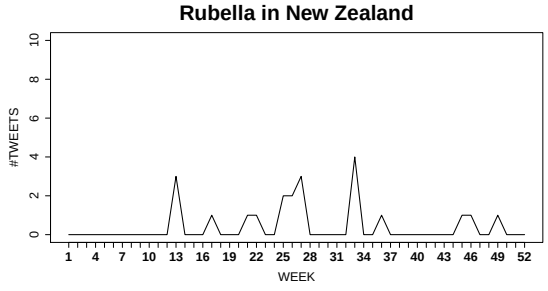

FIGURE 5.6: High Oscillation and Low Magnitude

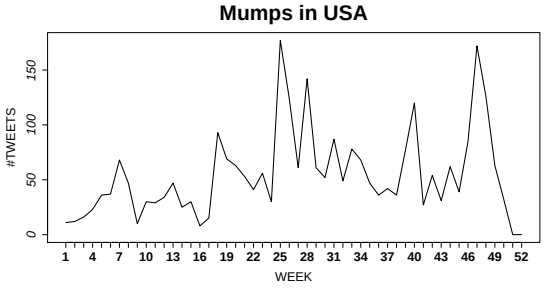

FIGURE 5.7: High Oscillation and High Magnitudes

A time series with a low magnitude, means that very few tweets can trigger a signal. This situation occurs in when are exploited only English tweets for outbreaks occurring in locations where English is not the main language or little international coverage is given to an event. Consider the outbreak "Rubella in New Zealand" (Figure 5.6), as an example of low magnitude time series, note that the maximum count is only 4 tweets

per day. It is expected that none of the detection methods would have major trouble generating a signal for this kind of pattern. Complex information filtering techniques are not required in this case given that the few tweets associated to the signals can be easily verified by the epidemiologists, and verify if indeed the tweets detected indicate a real outbreak.

A time series with low oscillation (Figure 5.4 and 5.5) indicate that the observed variable for the event noticeably peaks within the outbreaks period. This is the case of outbreaks such as "EHEC in Germany", "Botulism in Finland" or "Ebola in Uganda" - all of which are food-borne and cause high international and media coverage due to the ease with which they spread. Since the number of tweets is zero outside of the outbreak period, every algorithm will produce a signal for such outbreaks.

A time series with high oscillation and high magnitude (Figure 5.7) occurs when: 1) a disease occurs continuously in a country, such as mumps or leptospirosis; or 2) the disease is a highly ambiguous term, such as for anthrax. In these cases algorithm support is essential, since the amount of data is too large to survey manually, and it is not trivial to identify significant aberrations.

In the following, I discuss the results of the different biosurveillance algorithms. I recall that these results are related only to the most difficult class of outbreak-related time series data with high oscillation and high magnitude (outbreak IDs 2, 9, 10, and 11).

In Figure 5.8 are reported the averaged results in terms of Sensitivity, PPV and F-measure, for all the considered algorithms, using the best parameter setting identified by grid search.

From these results I can observe that:

1. In general, the results are not particularly positive, in terms of F-measure. FS and EWMA perform better with F-measure average value over 0.6. I consider that this is mainly due to the outbreak pattern of high oscillation and high magnitude.

2. For sudden outbreaks, such as the 2011 EHEC outbreak in Germany, it was obtained an F-measure of 0.95. Regardless of parameter settings, any of the detection methods will produce a signal for such kind of outbreaks [54].

3. Given the high variability of some outbreak patterns, there is no unique setting of the parameters able to deal with all considered diseases, but they should be tuned for each case based on historical data.

4. Note that even though the FA algorithm is widely used in existing biosurveillance systems, it performs worse than other algorithms in terms of F-measure. However,

it achieves a high PPV. Therefore, the choice of this algorithm strongly depends on the required tradeoff between precision and recall.

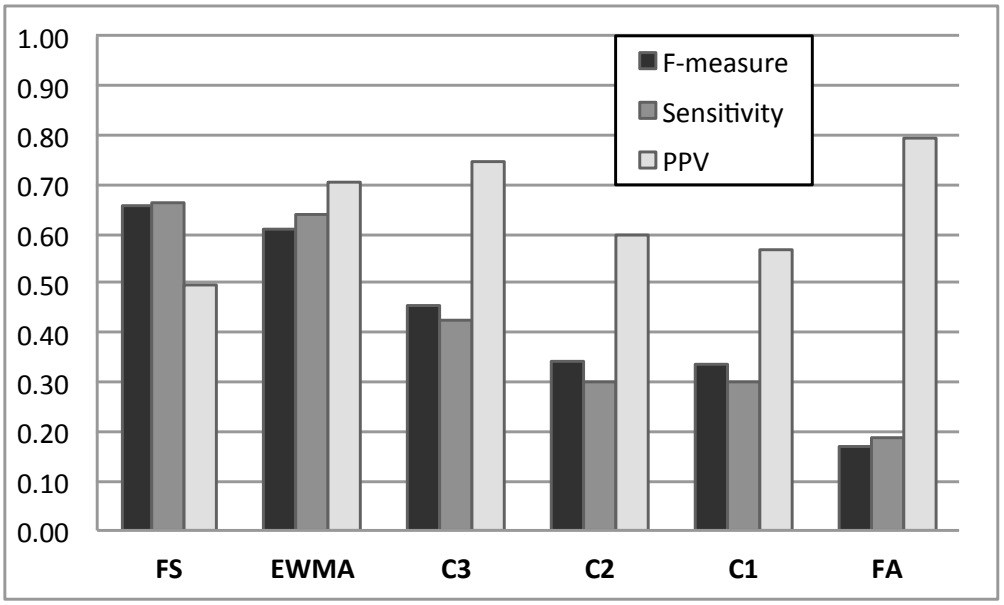

FIGURE 5.8: Performance of different algorithms measured by F-measure, Sensitivity and PPV.

5.3.3 Discussion

Research Question 3: *How external sources can be exploited for complementing traditional information?*

As mentioned earlier, the aim of this work is to detect outbreak events for general diseases that are not only seasonal, but also sporadic diseases that occur in low tweet-density regions. The results obtained show that events mentioned in Twitter could contribute towards generating early warning signals and inferring the existence of disease outbreak. However based on this analysis, it results that there are open challenges for automatically detecting disease outbreak from the stream. The first limitation of this study lies on the availability of standardized set of outbreaks and rich and abundant historical data. Having available information about outbreaks allows to compute more accurate analysis on time series data. In Section 5.3.1 I proposed a method for automatically extracting real-world events from unstructured text documents in order to build a ground truth of outbreak events from official reports that then could be compared and correlated with Twitter timeseries. In addition, I proposed a machine learning approach to determining the relevance of temporal expressions associated to a given event. The proposed features are based on annotated documents and domain-specific heuristics. Through experiments using real-world dataset, I showed that the proposed approach is able to identify relevant

temporal expressions for an event with good accuracy. In general the process of creating the ground truth for disease outbreaks requires information extraction techniques, namely, different NLP tools for extracting relevant information. Unfortunately, the accuracy of such tools are not nearly 100%, which has a severe impact to the coverage (number) and quality of outbreak ground truth found. For example, a place name are ambiguous and can be wrongly determined as the country of an outbreak as illustrated in this sentence *"The **Uganda** Virus Research confirms Ebola virus Sudan species"*.

The second limitation lies in the quality of Twitter data used in this study, for instance, I found many tweets to be either not-relevant or too sparse. Twitter data is highly ambiguous and noisy, which makes it difficult to detect an outbreak. A disease name mentioned in a tweet can have many contexts that are not relevant to an outbreak. In addition to ambiguity, using predefined disease names to gather data has limited the usefulness of the system. In fact, a Twitter user tends to use technical terms, e.g., "EHEC" or "E.Coli", when the term already has been used in a public or news media. One solution is to consider syndromic terms instead of medical conditions or disease names. In Section 5.3.2 I presented a study of outbreak events detection using different time series analysis methods. I identified four different classes of the detected events, based on the two characteristics: oscillation and magnitude. The only problematic type is the class of high oscillation and high magnitude time series. None of the investigated algorithms exhibits a high detection quality. In addition, a careful choice of parameters is required.

To conclude, gathering relevant tweets about a disease outbreak is not straightforward. Early signal generation could be improved for example applying sentiment analysis in order to classify tweets into news or personal health statuses. Event detection is only as good as entities detected. Moreover, regarding the assumption of the location of a tweet (i.e., text-containment location, geo-location, and a user's profile) each type should be treated/weighted differently for signal generation. Finally, moving beyond using only keywords/medical conditions, early outbreak detection can be performed by analyzing a group of people. e.g., any two or more clusters of social network users that are discussing outbreaks near in both time and location should be an indicator of an outbreak.

Chapter 6

Conclusions

In recent years several applications raised in order to support operators, working within different sectors, across the life cycle of a digital document, from its receipt until its processing and closure. What makes these instruments often unproductive is the fact that they were born to support the traditional, manual documentation process mainly based on the massive use of paper but, given the many technological and regulatory constraints, they can not completely replace the traditional paper process. Currently, there is a the strong coexistence of digital and paper-based information that makes the whole process expensive, unwieldy and partly fallacious. At the same time, during the creation of documents, there is a strong need to retrieve and reuse information that could be complementary to the contents handled. In this scenario it is necessary to find a way to manage information coming from different sources and that could be presented in different forms (multimedia, web pages, social media, microblogs) and languages (multilingualism).

In this thesis, I proposed a reconfigurable framework for knowledge management and document processing that has been instantiated, applied and tested in the medical domain. To this end it has been defined an architecture for heterogeneous and multi-language data management, in order to support the actors in the medical domain in accessing and retrieving useful information. In particular, this architecture supports the user in the medical record composition allowing to structuring, protect and organizing paper and digital medical record as well as managing heterogeneous information coming from the Web in order to identify outbreaks of epidemics.

6.1 Contribution

In this thesis I proposed an innovative framework that is based on a semantic methodology to transform unstructured data in a structured way. I illustrated its adoption in the medical domain to put in evidence the potentiality of the proposed approach from multiple perspectives.

The definition of the architecture for knowledge management and document processing made several contributions to the state of the art in information extraction, data mining and semantic document processing applied to heterogeneous and unstructured data.

- As regards to the *document classification* I proposed a methodology for automatic document categorization based on the adoption of unsupervised learning technique (clustering ensemble). I represented the documents in the vector space model by means of feature based on semantic processing. From this were built tree different classes of vector space models considering respectively terms, lemmas and synonyms (concepts). Those vector space models were adopted document clustering. I combined different results of X-means clustering algorithm executed with three different vectors space models which include syntactic and semantic content representation. For the experimental campaign it was adopted a corpus composed by real medical records that was digitalized by means of OCR. I then used the scanned data to perform the experiments. Since the dataset used is composed by digitalized by OCR medical records, the data used for testing present some noise. The noisy terms are discarded by means of the preprocessing phase of the semantic methodology and so the document are represented by the terms correctly recognized by the OCR procedure. It implies that the document representation in the vector space considers a subset of the original terms that occurs in the paper medical records. In the experiments I adopted the dataset in order to build the vector space model classes. For each classes, I built two kind of vector space model: in the first case I considered the dataset composed by the whole medical records; in the second case I considered only the first two pages of each medical record that report a summary of the whole medical record.

  The results showed that although the use of concepts allows us to make documents' partition by topic, it introduces noise, making the generated partitions worse than the ones obtained by using only Lemmas or Terms. On the other hand, the usage of semantic information combined with the syntactical ones allowed us to improve the obtained results.

- In the last years, the eHealth systems are considerably improving the quality and performance of services that an hospital is able to provide to its patients and his

workers. Up to date, many systems are based on document management systems and cannot benefit of new system design techniques to structure data and enforce fine-grain access control policies. Indeed, the medical records, especially the old ones, are just digitalized and made available to users. Being a monolithic resource, it is difficult to enforce proper security rules to guarantee privacy and confidentiality of data. In this thesis I have analyzed the security requirements of medical records and proposed a semantic approach to analyze the text, retrieve information from specific parts of the document that can be useful to classify them from a security point of view and, finally, associate a set of security rules that can be enforced on those parts. I have illustrated the adoption of the methodology on a simple case study to put in evidence the potentiality of the proposed methodology. The adoption of the semantic analysis on data is very promising and can strongly help in facing security issues that arise once data are made available for new potential applications.

- Through an empirical analysis, I conducted the first study on using tweets for detection **multiple** outbreaks (as opposed to previous work only studied one *single* medical condition, e.g., influenza or dengue) in *several* countries including those where the Internet is still lacking that are most at-risk for emerging and re-emerging diseases.

  I proposed an approach to automatically extracting outbreak events from unstructured raw documents by considering temporal and geographic expressions in a document. In addition, I proposed a machine learning approach to determining the relevance of temporal expressions associated to a given event. The proposed features are based on annotated documents and domain-specific heuristics. Through experiments using real-world dataset, I showed that the proposed approach is able to identify relevant temporal expressions for an event with good accuracy.

  The extracted events were used as *outbreak ground truth* for analyzing the timeliness of Twitter data by cross correlation, with respect to two time dimension associated to the outbreaks. In general the process of creating the ground truth for disease outbreaks requires information extraction techniques, namely, different NLP tools for extracting relevant information. Unfortunately, the accuracy of such tools are not nearly 100%, which has a severe impact to the coverage (number) and quality of outbreak ground truth found. For example, a place name are ambiguous and can be wrongly determined as the country of an outbreak as illustrated in this sentence *The **Uganda** Virus Research confirms Ebola virus Sudan species.*

  In order to detect outbreak events from twitter I used well known algorithms for burst detection used in biosurveillance. I set out to determine the ideal algorithms and parameters for categories of disease outbreaks. The results obtained

show that events mentioned in Twitter could contribute towards generating early warning signals and inferring the existence of disease outbreak. However based on this analysis, it results that there are open challenges for automatically detecting disease outbreak from the stream. The first limitation of this study lies on the availability of standardized set of outbreaks and rich and abundant historical data. Having available information about outbreaks allows to compute more accurate analysis on time series data. The second limitation lies in the quality of Twitter data used in this study, for instance, I found many tweets to be either not-relevant or too sparse. Twitter data is highly ambiguous and noisy, which makes it difficult to detect an outbreak. A disease name mentioned in a tweet can have many contexts that are not relevant to an outbreak. In addition to ambiguity, using predefined disease names to gather data has limited the usefulness of the system. In fact, a Twitter user tends to use technical terms, e.g., "EHEC" or "E.Coli", when the term already has been used in a public or news media. One solution is to consider syndromic terms instead of medical conditions or disease names. I presented a study of outbreak events detection using different time series analysis methods. I identified four different classes of the detected events, based on the two characteristics: oscillation and magnitude. The only problematic type is the class of high oscillation and high magnitude time series. None of the investigated algorithms exhibits a high detection quality. In addition, a careful choice of parameters is required.

To conclude, gathering relevant tweets about a disease outbreak is not straightforward. Early signal generation could be improved for example applying sentiment analysis in order to classify tweets into news or personal health statuses. Event detection is only as good as entities detected. Moreover, regarding the assumption of the location of a tweet (i.e., text-containment location, geo-location, and a user's profile) each type should be treated/weighted differently for signal generation. Finally, moving beyond using only keywords/medical conditions, early outbreak detection can be performed by analyzing a group of people. e.g., any two or more clusters of social network users that are discussing outbreaks near in both time and location should be an indicator of an outbreak.

The formalization of the general framework for data management is very promising and it is needed to investigate other methodologies and to integrate different analysis approaches with a close look at the huge number of new applications that can derive. There are still open issues that can be addressed as for example: consider more features and more clustering algorithms for the proposed document organization methodology; improving event extraction from raw documents identifying more features and apply

a machine learning technique for ranking temporal expressions in a document by relevance; investigate how timeseries patterns can be predicted by considering different characteristics as for example, for the context considered dealing with disease outbreaks, severe/common, easily spread/not (contagiousness).

Appendix A

The medical Record in HL7

A fragment of a Medical Record:

```
Ospedale Santo Bono - reparto Pronto Soccorso
Data 03/03/2005 ore 18,00

Paziente

Nome: Emma P.
Cognome: Esposito
Nata il: 09/03/1980

Diagnosi di entrata-la paziente  tranquilla e collaborante,
serena nell' espressione maxillo-facciale.
Somministrare per una settimana una compressa di asp309kz.
```

The resulting structure coded according to the HL7 standard:

```
<ExHL7 xmlns="urn:hl7-org:v3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="urn:hl7-org:v3 ExHL7.xsd">
 <id root="1.1" extension="batchID here" assigningAuthorityName="MessageSender"/>
 <creationTime value="20050303180027"/>
 <versionCode code="V3PR1"/>
 <interactionId root="1.1.6" extension="ExHL7" assigningAuthorityName="HL7"/>
 <receiver typeCode="RCV">
  <device classCode="DEV" determinerCode="INSTANCE">
   <id root="1.4.7"/>
    </device>
 </receiver>
 <sender typeCode="SND">
```

```
  <device classCode="DEV" determinerCode="INSTANCE">
   <id root="1.45.6"/>
  </device>
 </sender>
 <controlActProcess classCode="CACT" moodCode="EVN">
  <subject typeCode="SUBJ" contextConductionInd="false">
   <encounterEvent classCode="ENC" moodCode="EVN">
    <id root="1.56.3.4.7.5" extension="122345"
    assigningAuthorityName="SantoBono Pronto soccorso"/>
    <code code="EMER" codeSystem="2.16.840"/>
    <statusCode code="active"/>
    <subject contextControlCode="OP">
     <patient classCode="PAT">
      <id root="1.56" extension="55321" assigningAuthorityName="SantoBono"/>
      <patientPerson classCode="PSN" determinerCode="INSTANCE">
       <name>
        <given>Emma</given>
        <given>P</given>
        <family>Esposito</family>
       </name>
       <administrativeGenderCode code="F" codeSystem="2.16.840"/>
        <birthTime value="19800309"/>
      </patientPerson>
     </patient>
    </subject>
   </encounterEvent>
  </subject>
  <subject typeCode="SUBJ" contextConductionInd="false">
   <investigationEvent>
    <reaction>
     <!--Describe Event or Problem -->
      <text mediaType="text/plain">Diagnosi di entrata-la paziente   tranquilla
      e collaborante, serena nell' espressione maxillo-facciale</text>
    </reaction>
   </investigationEvent>
   <SubstanceAdministrationEvent>
    <id>asp309kz<\id>
    <text>somministare per una settimana<\text>
    <doseQuantity>1<\doseQuantity>
   </SubstanceAdministrationEvent>
</controlActProcess>
</ExHL7>
```

Bibliography

[1] Collins english dictionary - complete unabridged 10th edition. Mar 2013.

[2] Kjersti Aas and Line Eikvil. Text categorisation: A survey, 1999.

[3] Adobe. A primer on electronic document security, 2004.

[4] Eugene Agichtein and Luis Gravano. Snowball: extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, 2000.

[5] Bogdan Alexe, Mauricio A. Hernandez, Kirsten W. Hildrum, Rajasekar Krishnamurthy, Georgia Koutrika, Meenakshi Nagarajan, Haggai Roitman, Michal Shmueli-Scheuer, Ioana R. Stanoi, Chitra Venkatramani, and Rohit Wagle. Surfacing time-critical insights from social media. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2012.

[6] Alfresco. Alfresco Software. URL http://www.alfresco.com.

[7] Flora Amato. Methodologies and techniques for semantic management of documents in dematerialization processes. In *Ph.D. Thesis in Computer and Control Engineering*. University of Naples Federcico II, 2009.

[8] Flora Amato, Antonino Mazzeo, Antonio Penta, and Antonio Picariello. Using NLP and Ontologies for Notary Document Management Systems. In *Proceedings of the 19th International Conference on Database and Expert Systems Application*, 2008.

[9] Flora Amato, Antonino Mazzeo, Antonio Penta, and Antonio Picariello. Knowledge representation and management for e-government documents. In Antonino Mazzeo, Roberto Bellini, and Gianmario Motta, editors, *E-Government Ict Professionalism and Competences Service Science*, volume 280 of *IFIP International Federation for Information Processing*, pages 31–40. Springer Boston, 2008.

[10] Flora Amato, Antonino Mazzeo, Vincenzo Moscato, and Antonio Picariello. A system for semantic retrieval and long-term preservation of multimedia documents

in the e-government domain. *International Journal of Web Grid Services*, 5:323–338, December 2009.

[11] Flora Amato, Valentina Casola, Antonino Mazzeo, and Sara Romano. A semantic based methodology to classify and protect sensitive data in medical records. In *Sixth International Conference on Information Assurance and Security*, 2010.

[12] Flora Amato, Valentina Casola, Antonino Mazzeo, and Sara Romano. An innovative framework for securing unstructured documents. In lvaro Herrero and Emilio Corchado, editors, *Computational Intelligence in Security for Information Systems*, volume 6694 of *Lecture Notes in Computer Science*, pages 251–258. 2011.

[13] Flora Amato, Valentina Casola, Nicola Mazzocca, and Sara Romano. A semantic-based document processing framework: A security perspective. In *International Conference on Complex, Intelligent and Software Intensive Systems*, 2011.

[14] Flora Amato, Valentina Casola, Nicola Mazzocca, and Sara Romano. A semantic approach for fine-grain access control of e-health documents. *Logic Journal of IGPL*, 2012.

[15] Flora Amato, Valentina Casola, Sara Romano, and Antonino Mazzeo. A semantic based framework to identify and protect e-health critical resources. *Journal of Information Assurance and Security*, 7(4):296–306, 2012.

[16] Flora Amato, Francesco Gargiulo, Antonino Mazzeo, Sara Romano, and Carlo Sansone. Combining syntactic and semantic vector space models in the health domain by using a clustering ensemble. In *Proceedings of International Conference on Helth Informatics*, 2013.

[17] Douglas E. Appelt. Introduction to information extraction. *AI Commun.*, 12(3): 161–172, 1999.

[18] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2011.

[19] S.C. Bagui. Combining pattern classifiers: methods and algorithms. *Technometrics*, 47(4):517–518, 2005.

[20] Cataldo Basile, Antonio Lioy, Marco Vallini, and Salvatore Scozzi. Ontology-based security policy translation. *Journal of Information Assurance and Security*, 5:437–445, 2010.

[21] Michele Basseville and Igor V. Nikiforov. *Detection of Abrupt changes: Theory and Application*. Prentice Hall, 1993.

[22] Thomas Beale and Sam Heard. The openEHR EHR Service Model. *Revision 02 openEHR Reference Model the openEHR foundation*, 2003.

[23] Thomas Beale and Sam Heard. Archetype definitions and principles. *Revision 06 March*, pages 1–15, 2007.

[24] Moritz Y. Becker and Peter Sewell. Cassandra: flexible trust management, applied to electronic health records. In *17th IEEE Computer Security Foundations Workshop*, 2004.

[25] D. Elliot Bell and Leonard J. LaPadula. Secure computer systems: Mathematical foundations and model, 1973.

[26] Klaus Berberich, Srikanta Bedathur, Omar Alonso, and Gerhard Weikum. A language modeling approach for temporal information needs. In *Proceedings of European Conference on Information Retrieval*, 2010.

[27] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, 2002.

[28] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. 2001.

[29] Tim Berners-Lee, Wendy Hall, and James A Hendler. *A framework for web science*. Now Pub, 2006.

[30] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. KNIME: The Konstanz Information Miner. In Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, and Reinhold Decker, editors, *Data Analysis, Machine Learning and Applications*, chapter 38, pages 319–326. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-78239-1. doi: 10.1007/978-3-540-78246-9_38. URL http://dx.doi.org/10.1007/978-3-540-78246-9_38.

[31] Rafael Bhatti, Khalid Moidu, and Arif Ghafoor. Policy-based security management for federated healthcare databases (or rhios). In *Proceedings of the international workshop on Healthcare Information and Knowledge Management*, 2006.

[32] Douglas Biber, Randi Reppen, and Susan Conrad. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, 1998.

[33] Susanne Briet. *Qu'est-ce que la documentation*. Paris: EDIT, 1951.

[34] Michael K. Buckland. What is a "document"? *Journal of the American Society for Information Science (JASIS)*, 48(9):804–809, 1997.

[35] Howard Burkom. Accessible alerting algorithms for biosurveillance. In *National Syndromic Surveillance Conference 2005*, 2005.

[36] Christopher S. Butler. *Statistics in Linguistics*. Blackwell, Oxford, 1985.

[37] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, 2010.

[38] Jill Cheng, Melissa Cline, John Martin, David Finkelstein, Tarif Awad, David Kulp, and Michael A Siani-Rose. A knowledge-based clustering algorithm driven by gene ontology. *Journal of Biopharmaceutical Statistics*, 14(3):687–700, 2004.

[39] Nigel Collier. What's Unusual in Online Disease Outbreak News? *Journal of Biomedical Semantics*, 1(1):2, 2010.

[40] Nigel Collier and Son Doan. Syndromic classification of twitter messages. *Computing Research Repository (CoRR)*, 2011.

[41] Nigel Collier, Nguyen Truong Son, and Ngoc Mai Nguyen. OMG U got flu? Analysis of shared health messages for bio-surveillance. *Computing Research Repository (CoRR)*, 2011.

[42] Dublin Core. Doublin core metadata initiative, 2004. URL http://dublincore.org/.

[43] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, 2010.

[44] Ido Dagan and Ken Church. Termight: identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied natural language processing*, 1994.

[45] Na Dai, Milad Shokouhi, and Brian D. Davison. Learning to rank for freshness and relevance. In *Proceeding of Special Interest Group on Information Retrieval*, 2011.

[46] David L. Davies and Donald W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, April 1979.

[47] John Davies, Marko Grobelnik, and Dunja Mladenic. *Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[48] John Davies, Marko Grobelnik, and Dunja Mladenic. Challenges of semantic knowledge management. In John Davies, Marko Grobelnik, and Dunja Mladenic, editors, *Semantic Knowledge Management*, pages 245–247. Springer Berlin Heidelberg, 2009.

[49] Sergio Decherchi, Paolo Gastaldo, Judith Redi, and Rodolfo Zunino. A text clustering framework for information retrieval. *Journal of Information Assurance and Security*, 4:174–182, 2009.

[50] Gianluca Demartini, Malik Muhammad Saad Missen, Roi Blanco, and Hugo Zaragoza. Taer: time-aware entity retrieval exploiting the past to find relevant entities in news articles. In *Proceedings of ACM International Conference on Information and Knowledge Management*, 2010.

[51] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[52] Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Mach. Learn.*, 42:143–175, January 2001.

[53] Fernando Diaz and Rosie Jones. Using temporal profiles of queries for precision prediction. In *Proceedings of ACM SIGIR Special Interest Group on Information Retrieval*, 2004.

[54] Ernesto Diaz-Aviles, Avaré Stewart, Edward Velasco, Kerstin Denecke, and Wolfgang Nejdl. Epidemic Intelligence for the Crowd, by the Crowd. In *International AAAI Conference on Weblogs and Social Media*, 2012.

[55] DICOM. Health informatics digital imaging and communication in medicine (DICOM), 2006. URL http://www.hl7.org/documentcenter/.

[56] DICOM. DICOM SR Basic Diagnostic Imaging Report to HL7, 2006. URL http://www.hl7.org/documentcenter/.

[57] Robert H. Dolin, Liora Alschuler, Sandy Boyer, Calvin Beebe, Fred M. Behlen, Paul V. Biron, and Amnon Shabo Shvo. Hl7 clinical document architecture, release 2. *Journal of the American Medical Informatics Association*, 13(1):30–39, 2006.

[58] Carlotta Domeniconi and Muna Al-Razgan. Weighted cluster ensembles: Methods and analysis. *ACM Transactions on Knowledge Discovery from Data*, 2(4):17:1–17:40, 2009.

[59] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th international conference on World wide web*, New York, NY, USA, 2010.

[60] J. C. Dunn. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[61] J. Durkin. *Expert systems: design and development.* Macmillan, 1994.

[62] eDocXL. eDocXL Document Management Software. URL http://www.edoc.gr.

[63] Jonathan L. Elsas and Susan T. Dumais. Leveraging temporal dynamics of document content in relevance ranking. In *Proceedings of International Conference of Web Search and Data Mining*, 2010.

[64] Michael Emch, Caryl Feldacker, M. Sirajul Islam, and Mohammad Ali. Seasonality of cholera from 1974 to 2005: a review of global patterns. *International Journal of Health Geographics*, 7(1), 2008.

[65] Empolis. Empolis Information Management. URL http://www.empolis.com.

[66] David M. Eyers, Jean Bacon, and Ken Moody. Oasis role-based access control for electronic health records. *IEE Proceedings - Software*, 153(1):16–23, 2006.

[67] Tamburini Fabio. Annotazione grammaticale e lemmatizzazione di corpora in italiano, 2000.

[68] Lu Fan, William J Buchanan, Christoph Thuemmler, Owen Lo, Abou Sofyane Khedim, Omair Uthmani, Alistair Lawson, and Derek Bell. DACAR platform for eHealth services cloud. In *Proceedings of the 4th International Conference on Cloud Computing*, 2011.

[69] C. P. Farrington, N. J. Andrews, A. D. Beale, and M. A. Catchpole. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(3):pp. 547–563, 1996.

[70] Edward A Feigenbaum and Pamela McCorduck. The fifth generation: artificial intelligence and japan's computer challenge to the world. 1983.

[71] Samah Jamal Fodeh, William F Punch, and Pang-Ning Tan. Combining statistics and semantics via ensemble model for document clustering. In *Proceedings of ACM Symposium on Applied Computing*, 2009.

[72] P. Foggia, G. Percannella, C. Sansone, and M. Vento. Benchmarking graph-based clustering algorithms. *Image Vision Compututing*, 27:979–988, 2009.

[73] Edgar Gonzàlez and Jordi Turmo. Comparing non-parametric ensemble methods for document clustering. In *Proceedings of International conference on Applications of Natural Language Processing to Information Systems*, 2008.

[74] Ralph Grishman. Information extraction: techniques and challenges. In *Information Extraction (International Summer School SCIE-97)*, 1997.

[75] Ralph Grishman. Information extraction: Capabilities and challenges, 2012.

[76] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5:199–220, 1993.

[77] Khaled M. Hammouda and Mohamed S. Kamel. Document similarity using a phrase indexing graph model. *Knowledge and Information Systems*, 6:710–727, November 2004. ISSN 0219-1377.

[78] James Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, and Daniel Weitzner. Web science: an interdisciplinary approach to understanding the web. *Communictions of the ACM*, 51(7):60–69, July 2008.

[79] Jerry R. Hobbs and Ellen Riloff. Information extraction. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010.

[80] Johannes Hoffart, Fabian Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence Journal, Special Issue on Wikipedia and Semi-Structured Resources*, 2012.

[81] Anna Huang, David Milne, Eibe Frank, and Ian Witten. Clustering documents using a wikipedia-based concept representation. In Thanaruk Theeramunkong, Boonserm Kijsirikul, Nick Cercone, and Tu-Bao Ho, editors, *Advances in Knowledge Discovery and Data Mining*, volume 5476 of *Lecture Notes in Computer Science*, pages 628–636. 2009.

[82] Lori Hutwagner, William Thompson, G Matthew Seeman, and Tracee Treadwell. The bioterrorism preparedness and response Early Aberration Reporting System (EARS). *Journal of urban health bulletin of the New York Academy of Medicine*, 80(2 Suppl 1):i89–i96, 2003.

[83] Ilias Iakovidis. Towards personal health record: current situation, obstacles and trends in implementation of electronic healthcare record in europe. *International Journal of Medical Informatics*, 52(1-3):105–115, October 1998.

[84] Rituraj Jain. Improvement in software development process and software product through knowledge management. *International Journal of Computer Technology and Applications*, 02(05):1557–1562, 2011.

[85] Adam Jatowt, Yukiko Kawai, and Katsumi Tanaka. Temporal ranking of search engine results. In *Proceedings of International Conference on Web Information Systems Engineering*, 2005.

[86] Jing Jin, Gail-Joon Ahn, Hongxin Hu, Michael J. Covington, and Xinwen Zhang. Patient-centric authorization framework for electronic healthcare services. *Computers and Security*, 30(2-3):116 – 127, 2011. Special Issue on Access Control Methods and Technologies.

[87] Nattiya Kanhabua, Roi Blanco, and Michael Matthews. Ranking related news predictions. In *Proceeding of Special Interest Group on Information Retrieval*, 2011.

[88] Nattiya Kanhabua, Sara Romano, and Avaré Stewart. Identifying relevant temporal expressions for real-world events. In *SIGIR Workshop on Time-aware Information Access*, 2012.

[89] Nattiya Kanhabua, Sara Romano, Avaré Stewart, and Wolfgang Nejdl. Supporting temporal analytics for health-related events in microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012.

[90] Graeme D. Kennedy. *An introduction to corpus linguistics*. Longman, 1998.

[91] KeyMark. KeyMark Document Management Software. URL http://www.keymarkinc.com.

[92] Sharib A. Khan. Handbook of Biosurveillance. *Journal of Biomedical Informatics*, 2007.

[93] Alison Kidd. The marks are on the knowledge worker. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1994.

[94] Patrick Kierkegaard. Electronic health record: Wiring europes healthcare. *Computer Law & Security Review*, 27(5):503–515, 2011.

[95] Jun-Tae Kim and Dan I. Moldovan. Acquisition of linguistic patterns for knowledge-based information extraction. *IEEE Transactions on Knowledge and Data Engineering*, 7(5):713–724, October 1995.

[96] Aas Kjersti and Eikvil Line. Text categorisation: A survey., 1999.

[97] Anagha Kulkarni, Jaime Teevan, Krysta M. Svore, and Susan T. Dumais. Understanding temporal query dynamics. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011.

[98] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, 2004.

[99] Vasileios Lampos and Nello Cristianini. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology*, 3, 2011.

[100] G. N. Lance and W. T. Williams. A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems. *The Computer Journal*, 9(4):373–380, February 1967.

[101] Xiaoyan Li and W. Bruce Croft. Time-based language models. In *Proceedings of International Conference on Information and Knowledge Management*, 2003.

[102] Donald Lindberg, Betsy Humphreys, and Alexa McCray. The unified medical language system. *Methods of Information in Medicine*, 32(4):281–291, 1993.

[103] John Locke. *Of Knowledge and Probability*. An Essay: Concerning Human Understanding. BOOK IV. Dover, 1689.

[104] ManagePoint. ManagePoint Document Management. URL http://www.managepoint.com.au.

[105] Christopher D. Manning and Hinrich Schuetze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1 edition, June 1999.

[106] Frank Manola and Eric Miller. RDF Primer, 2004. URL http://www.w3.org/TR/rdf-primer/.

[107] Donald Metzler, Rosie Jones, Fuchun Peng, and Ruiqiang Zhang. Improving search relevance for implicitly temporal queries. In *Proceedings of Special Interest Group on Information Retrieval*, 2009.

[108] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244, 1990.

[109] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Introduction to WordNet: An On-line Lexical Database. Technical report, Cognitive Science Laboratory, Princeton University, 993.

[110] EHR Model. Health informatics - Electronic Health Record Communication – Part 1: Reference Model, 2008.

[111] Marie Francine Moens. *Information Extraction: Algorithms and Prospects in a Retrieval Context*. The Springer international series on information retrieval. Springer, 2006.

[112] OpenCalais. OpenCalais, http://www.opencalais.com/.

[113] OpenDocMan. Open Source Document Management System. URL http://www.opendocman.com.

[114] OpenEHR. OpenEHR Community, 2005. URL http://www.openehr.org/.

[115] OpenKM. OpenKM Knowledge Management. URL http://www.openkm.com.

[116] OpenNLP. OpenNLP, http://opennlp.apache.org/.

[117] Patrick Pantel and Marco Pennacchiotti. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 2006.

[118] Michael J. Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *Proceedings of International AAAI Conference on Weblogs and Social Media*, 2011.

[119] Roger T. Pdauque. Document: Form, sign and medium, as reformulated for electronic documents. *STIC-CNRS*, 2003.

[120] Dau Pelleg and Andrew Moore. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *In Proceedings of the 17th International Conf. on Machine Learning*, 2000.

[121] ProMEDMail. ProMED-mail, http://www.promedmail.org/.

[122] Eric Prudhommeaux and Andy Seaborne. SPARQL Query Language for RDF, 2005. URL http://www.w3.org/TR/rdf-sparql-query/.

[123] QuadraMed. QuadraMed Quality Care. Financial Health. URL http://www.quadramed.com/.

[124] Greg Ridgeway, David Madigan, and Thomas Richardson. Interpretable boosted naive bayes classification. In *4th International Conference on Knowledge Discovery and Data Mining*, 1998.

[125] Ellen Riloff. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the eleventh national conference on Artificial intelligence*, 1993.

[126] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of Intenational World Wide Web Conference*, 2010.

[127] Phillipe Salembier and Thomas Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc., New York, NY, USA, 2002.

[128] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. In *Information processing and management*, pages 513–523, 1988.

[129] Gerard Salton, Andrew Wong, and ChungShu Yang. A vector space model for automatic indexing. *Communications ACM*, 18:613–620, November 1975.

[130] Eric Sanjuan and Fidelia Ibekwe-sanjuan. Phrase clustering without document context, 2006.

[131] Bolasco Sergio. *Statistica testuale e text mining: alcuni paradigmi applicativi.* Quaderni di Statistica, 2005.

[132] Hiroyuki Shinnou and Minoru Sasaki. Ensemble document clustering using weighted hypergraph generated by NMF. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 2007.

[133] Milad Shokouhi. Detecting seasonal queries by time-series analysis. In *Proceeding of Special Interest Group on Information Retrieval*, 2011.

[134] MiguelbAngel Sicilia. Metadata, semantics, and ontology: providing meaning to information resources. *International Journal of Metadata, Semantics and Ontologies*, 1(1):83–86, January 2006.

[135] Alexander V. Smirnov, Mikhail Pashkin, Nikolai Chilov, Tatiana Levashova, Andrew Krizhanovsky, and Alexey Kashevnik. Ontology-based users and requests clustering in customer service management system. *Computing Research Repository (CoRR)*, pages –1–1, 2005.

[136] Stephen Soderland, David Fisher, Jonathan Aseltine, and Wendy Lehnert. CRYSTAL inducing a conceptual dictionary. In *Proceedings of the 14th international joint conference on Artificial intelligence*, 1995.

[137] Mustafa Sofean, Avaré Stewart, Matthew Smith, and Kerstin Denecke. Medical case-driven classification of microblogs: Characteristics and annotation. In *Proceedings of SIGHIT International Health Informatics Symposium*, 2012.

[138] SoftTech. SoftTech Document Management Software. URL http://www.softtechhealth.com.

[139] Jannik Strötgen and Michael Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 2010.

[140] Jannik Strötgen, Michael Gertz, and Conny Junghans. An event-centric model for multilingual document similarity. In *Proceeding of Special Interest Group on Information Retrieval*, 2011.

[141] Jannik Strötgen, Omar Alonso, and Michael Gertz. Identification of top relevant temporal expressions in documents. In *Proceeding of the 2nd Temporal Web Analytics Workshop*, 2012.

[142] Martin Szomszor, Patty Kostkova, and Ed de Quincey. #swineflu: Twitter predicts swine flu outbreak in 2009. In *Proceedings of eHealth*, 2010.

[143] IHE technical Frameworks. Integrating the Healthcare Enterprise IT Infrastructure Technical Framework, 2007.

[144] Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. #twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011.

[145] Sandro Vega-Pons and Jose Ruiz-Shulcloper. A Survey of Clustering Ensemble Algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(3):337–372, 2011.

[146] Marc Verhagen, Inderjeet Mani, Roser Sauri, Jessica Littman, Robert Knippen, Seok Bae Jang, Anna Rumshisky, John Phillips, and James Pustejovsky. Automating Temporal Annotation with TARSQI. In *Proceedings of Association for Computational Linguistics*, 2005.

[147] Manish Verma. Xml security: Control information access with xacml. *IBM developer Works*, 2004.

[148] Max Vlkel. From documents to knowledge models. In *Proceedings of the 4th Conference on Professional Knowledge Management*, 2007.

[149] Raphael Volz, Siegfried Handschuh, Steffen Staab, Ljiljana Stojanovic, and Nenad Stojanovic. Unveiling the hidden bride: deep annotation for mapping and migrating legacy data to the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1(2):187–206, 2004.

[150] Weka. Weka: Data Mining Software in Java, http://weka.sourceforge.net.

[151] WHOReport. WHO disease outbreak reports, http://www.who.int/csr/don/en/.

[152] Wikipedia. Wikipedia, the free encyclopedia, 2004. URL http://en.wikipedia.org/.

[153] WinDream. WinDream Document Management. URL http://www.windream.com.

[154] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann, 2005.

[155] World Wide Web Consortium (W3C). Extensible Markup Language (XML) 1.1, W3C Recommendation, 2004. URL http://www.w3.org/TR/xml11.

[156] Pingpeng Yuan, Yuqin Chen, Hai Jin, and Li Huang. MSVM-kNN: Combining SVM and k-NN for Multi-class Text Classification. *IEEE International Workshop on Semantic Computing and Systems,*, 2008.