



UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II
Dottorato di Ricerca in Ingegneria Informatica ed Automatica



Methodologies and Techniques for Semantic Management of Documents in Dematerialization Processes

FLORA AMATO

Tesi di Dottorato di Ricerca

(XXII Ciclo)

Novembre 2009

Tutor

Prof. Antonino Mazzeo

Coordinatore del Dottorato

Prof. Luigi P. Cordella

Dipartimento di Informatica e Sistemistica

A Francesco

La mia roccia

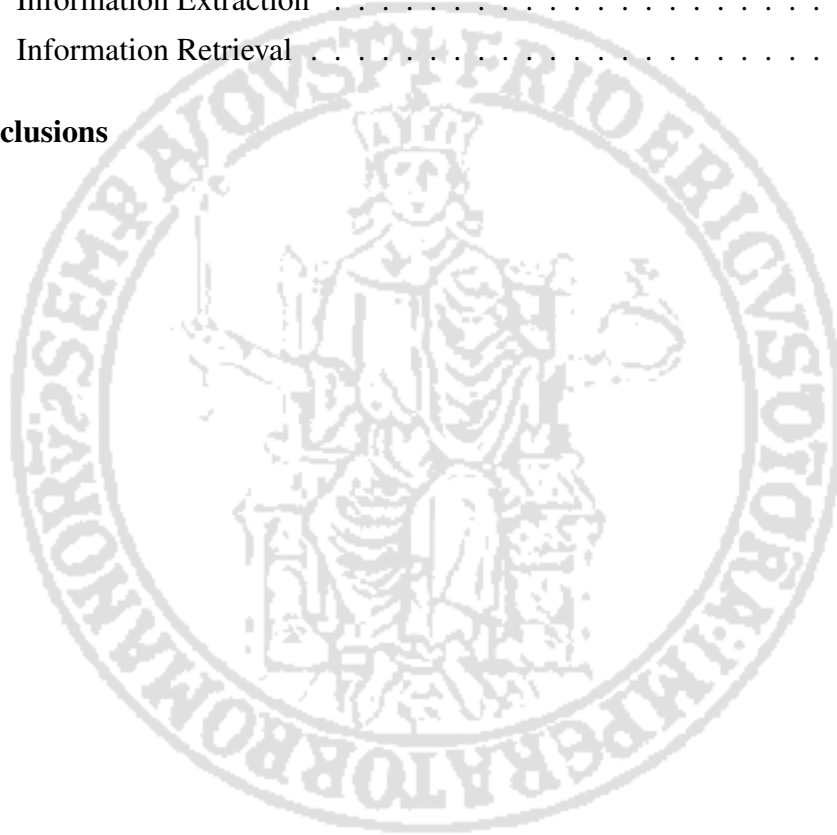
Nelle acque quiete e nel mare in tempesta

Contents

1	Semantic document processing: an Introduction	6
1.1	E-Government and Dematerialization Activity: Context and Open Issues	7
1.1.1	The Italian Perspective	9
1.2	Objectives	11
1.3	Motivating Examples	15
1.3.1	Juridical Documents	16
1.3.2	Practices of Telemedicine	16
1.4	Outline	18
2	Domain Characterization: Methodologies and State of the Art	19
2.1	Introduction to Domain Modeling	19
2.2	Ontology Definitions	21
2.3	Problems of interpretation in non-specialized domains: the ambiguity of natural language	25
2.4	Specialist languages and specialized domains	33
2.5	Concepts and Relations detection	37
2.5.1	Relevant Terms Recognition	39
	Text Preprocessing	41
	Morpho-syntactic analysis	44
2.5.2	Extraction of the relevant terminology	53
	Endogenous resources: the computation of the TFIDF index	54
	Exogenous resources: lexical comparisons	56
2.5.3	Identification of lexical-semantic relations	60

Creation of synsets: synonyms	63
3 State of the art in Document Management Systems	72
3.1 Document Management Systems	72
3.2 Multimedia Document Management Systems	76
3.2.1 Image Database Systems	76
3.2.2 Text Database Systems	78
3.3 Domain-Document Association	79
3.3.1 Feature Selection	79
3.4 Ontology driven human assisted Annotation	81
3.5 Information Retrieval	84
4 A Digital Document Model	88
4.1 A model of document suitable for e-government activity	88
4.1.1 Preliminary Definitions	91
4.1.2 E-Government Document Definition	92
4.2 The RDF Digital Document Model	94
5 An Architecture for Semantic Document Management System	98
5.1 General Process Overview	98
5.2 The System Architecture	101
5.2.1 The Text Processing Module	104
5.2.2 The Multimedia Processing Module	105
5.2.3 The Integration and Presentation modules	107
5.2.4 Document Processing	108
5.3 Domain Characterization	109
5.4 Domain-Document Association	120
5.4.1 Document Segmentation	120
5.4.2 Document Classification	124
5.5 Formal Information Structuring	128
5.6 Information Extraction and Ontology Population	135
5.7 Information Retrieval	140

5.8	System Description	142
5.8.1	Implementation Issues	149
6	Experimental Results for processing of documents in juridical domain	152
6.1	The Selected Corpus	152
6.2	Domain characterization: Relevant Terms Extraction	153
6.3	Documents Classification	154
6.4	Information Extraction	156
6.5	Information Retrieval	158
7	Conclusions	161



List of Figures

1.1	The Italian legislative Scenario	10
1.2	An Example of Italian Juridical Document	17
2.1	Domain Knowledge Formalization	26
2.2	Example of syntactic ambiguity	29
2.3	Intersections among language varieties	57
2.4	general process for extraction of Relevant Terminology	59
3.1	State of the Art in Commercial Document Management System . .	73
3.2	Example of scissed weak link	80
4.1	The Document Model	90
4.2	Digital Document RDF Model	97
5.1	General Schema of Whole Documents Processing	99
5.2	System Architecture	103
5.3	Lexical resources to seek legal terms	115
5.4	Luhn's law	117
5.5	Iterative Processing for identification of Peculiar Lexicon	118
5.6	Application of tree Partition Criteria on the same Act fragment . .	123
5.7	Association Document Segments ↔ Ontological Fragments . . .	125
5.8	Detailed workflow for features extraction	127
5.9	Detailed workflow for Segments Classification	129
5.10	<i>RDF-Extractor</i> (<i>RDFex</i>) algorithm	136
5.11	Example of Ruled Based Semantic Annotation	137

5.12	An Example of semantic annotation for a Notary Act	138
5.13	A section of <i>RDF</i> graph extracted from a Notary Act	139
5.14	a snapshot of information retrieval procedure	143
5.15	System Architecture	145
5.16	Interface for Information Extraction	149
6.1	the selected corpus	153
6.2	List Fragment of Extracted Relevant Terms	154
6.3	Example of Extensional Ontology Fragment	155
6.4	Classification Results	156
6.5	<i>TK</i> sets selection	159
6.6	Experimental Result for Information Retrieval task for selected categories	160

List of Tables

2.1	Example of word category ambiguity for the Italian language . . .	27
2.2	Examples of hyponyms and hyperonyms	69
2.3	Examples of hyponymic relations extracted from a corpus of legal documents	71
5.1	Chi-squared distance among the documents, the corpus and the peculiar lexical items	119
5.2	Cover rates of each document, the corpus and lexical peculiar index	119
6.1	Indexing Precision	157
6.2	Indexing Times	158

Acknowledgements

Voglio ringraziare innanzitutto il mio Tutor, il Prof. Antonino Mazzeo, che ha creduto in me. Grazie per tutte le volte che ha condiviso le sue idee, spesso assolutamente sorprendenti, e grazie anche per le volte in cui le ha cambiate.

Grazie per avermi guidato in contesti e compiti difficili, e grazie soprattutto per quando ha lasciato guidare me.

Grazie infine di avermi indirizzato in questi tre anni, intensi e difficili, ma al contempo bellissimi.

Ringrazio il Prof. Antonio Picariello, la Prof. Valentina Casola e il Prof. Vincenzo Moscato, che hanno voluto condividere con me un pezzo del loro percorso di ricerca e sono stati prima che coautori, la mia famiglia.

Ringrazio il Prof. Nicola Mazzocca, il Prof. Carlo Sansone, il Prof. Beniamino Di Martino, il Prof. Porfirio Tramontana e la Prof. Valeria Vittorini per il supporto che mi hanno dato durante tutto il lavoro che ha condotto alla tesi.

Grazie a Rosanna, che sempre discreta, mi ha fatto avvicinare al lato umanistico della semantica, a Sara, Tiziana, Francesco, Emanuela, Carmelo, Giuseppe, Andrea, Christian e Antonio con cui abbiamo condiviso progetti, preoccupazioni e risate.

Grazie alle mie sorelline, sempre meravigliosamente in grado di strapparmi un sorriso in ogni situazione, a mia madre per esserci stata davvero sempre, e a mio padre che mi ha supportato senza chiedermi perché mai avessi smesso di lavorare.

Concludo ringraziando Francesco, a cui dedico questo lavoro di tesi, per essere il mio compagno, il mio amico e il mio amore.

Napoli, Italy, 30 Novembre 2009

Flora

Knowledge management has become a challenging issue for almost all the e-Government based applications. One of the main issues for E-Government activities is to manage the great amount of available data efficiently.

The presence of a huge amount of information, in fact, is typical of bureaucratic processes, like the ones pertaining to public administrations. Such information is often recorded on paper or on different digital files and its management is very expensive, both in terms of space used for storing and in terms of time spent in searching for the documents of interest. Furthermore, the manual management of these documents is absolutely not error-free.

In order to efficiently access the information embedded in very large document repositories, techniques for semantic document management are required. They ensure a large and intense process of dematerialization and aim at eliminating or at least reducing, the amount of paper documents.

E-Government based applications need proper data models for information content characterization, in order to automatically transform unstructured (or sometimes semi-structured) documents into formally structured records, suitable for machine processing. Furthermore a way for presenting information contained in documents, depending on access policies and available technologies has to be provided. Finally different kinds of media elements, contained in digital documents, have to be managed. Indeed, nowadays, almost all the novel bureaucratic processes are characterized by both text and multimedia data (e. g. audio, still images, sometimes videos), which need to be properly handled, stored and distributed.

In this thesis, we present a novel model of digital documents for improving the dematerialization effectiveness, that constitutes the starting point for an information system able to manage documents streams efficiently. This model takes into account E-Government applications needs like as the respect of provisions in force and the adaptability to evolving technologies. At the best of our knowledge, the proposed model is one of the first attempts to give a single and unified characterization for the management of multimedia documents, pertaining to a bureaucratic domain as the E-Government one, on which a system of semantic procedures are

defined for the transformation of the non structured documents (pertaining to specialized domain) into structured data.

Furthermore, architecture for the management of the document whole life cycle has been proposed, which provides advanced functionalities for semantic processing, such as giving formal structure to document informative content, information extraction, semantic retrieval, indexing, storage, presentation, together with long-term preservation.



Chapter 1

Semantic document processing: an Introduction

E-government processes are dedicated to the improvement of the efficiency, expensiveness and accessibility of public administration services: dematerialization activities, introduced for properly managing bureaucratic documents, are among the main tasks of the E-Government works.

It is widely agreed that Semantic-based dematerialization process will greatly enhance systems and application procedures designed for e-Gov activity [4],[7],[8].

The dematerialization process implies the application of syntactic-semantic methodologies in order to automatically transform the unstructured or sometimes semi-structured document into formally structured records, suitable for machine treatment.

The core aspect related to a novel and efficient dematerialization process is the idea standing beyond the common document concept, that can be defined according to the Italian civil law[9], as the representation of acts, facts and figures directly made or by means of electronic processing, and stored on a intelligible support.

In other words, a document consists of objects such as text, images, drawings, structured data, operational codes, programs and movies, that, according to their relative position on the support, determine the shape and, consequently the structure of the document itself through the relationships between them.

During the various e-Government processing phases, that are really different

from an application domain to another, a document is processed and eventually stored on various kinds of media, properly defined in order to archive and preserve papers, photographic films and microfilms, VHS cassettes, Magnetic Tapes, DVD disks, and more.

In order to manage documents properly, Document Management System (DMS) are used. They were introduced in the early 1970 for converting paper documents into electronic images stored in computers. Once digitally captured, DMS allow for documents retrieval effortlessly and for sharing and accessing by multiple users. Nowadays DMS are becoming the fundamentals of most business information systems, giving user control over company knowledge, providing efficient retrieval and desktop integration, reducing error rates in documents manipulation and thus improving business performance.

With the use of standards for knowledge representation, DMS are evolving, from search engine, toward system able to integrate semantic search procedures into companies business processes. Such systems, however, are limited to provide additional semantic functionalities to existent document management features. At the best of our knowledge, there are no systems modeling multimedia document contents from a semantic perspective, thus providing a fully automated semantic management for them. Such process aims to structure the input documents and to allow for automatic extraction of targeted information, depending on formal representation of the domain associated to the documents, defined in a semi-automatic way, starting from the processable document themselves.

1.1 E-Government and Dematerialization Activity: Context and Open Issues

E-Government related activities involve the electronic management of public services, or processes of Governance. It concerns the reorganization of the bureaucratic processes in both central and local Public Administrations (PA). One of the main goals of e-Government is providing automatic management of documental flows, in order to optimize the work of the governmental offices and to offer to

users (citizens and businesses) faster, effective and accessible services.

E-Government can be considered as the application of Information and Communication Technology (ICT) to problems that typically belong to both Public Administration and legal domains.

The use of ICT in public administrations has been introduced some decades ago with a number of ad hoc projects, aiming at the automation of parts of information processing activities and integration of pre-existent legacy applications, devoted to the automation of the entire bureaucratic process.

Many initiatives, often supported by facilitated finances, were introduced in the eighties within the European Community in order to introduce the use of information systems in the PA, with the objective of supporting the principal bureaucratic processes within specific domains (ministries, local bodies, regions, etc.).

In the nineties and until the beginning of the present decade, with the spread of the Internet and the related technologies, the focus has been moved towards the opening of such systems to the web, in order to carry out initiatives of e-Government and define a first level of interconnection among administrations from different domains, principally in national, but also in international environment.

Nowadays, the process of combining the effectiveness of services and their transparency within the Public Administration context, goes through a strong automation of the internal processes involving the use of open systems, able to cooperate at application levels, following federate models devoted to perform inter-domain bureaucratic processing.

Almost all the e-Gov applications have dematerialization activities as a common and fundamental factor: information, previously stored using graphic marks on material (paper) supports, is made immaterial using a codified electronic representation, and can be nowadays stored on several digital supports such as memories, magnetic or optical disks, tapes or other mature technologies nowadays in use. Dematerialization is not only a normative and technological challenge but also an organizational matter involving various human resources. Transforming a bureaucratic organization based on paper into one based on electronic documents

is not easily achievable for complex entities as Public Administrations.

So far, we have described the main characteristic of the e-Gov system, in particular, we note that e-Gov processes are usually characterized by a large quantity of paper documents that need to be properly managed, stored and distributed. In order to reduce the amount of hard papers and to optimize information communication in terms of time and resources, it is widely agreed that a semantic-based dematerialization process will greatly enhance e-Gov systems and application procedures, improving the quality of services, enabling the diffusion and the access of the information of interest to all the authorized users in an efficient and transparent way.

The dematerialization process requires the application of syntactic-semantic methodologies in order to automatically transform the unstructured or sometimes semi-structured document into a formally structured, machine readable records. In this way, advanced functionalities for data management are provided, including the extraction of the relevant information [32], [4], the information representation according to the formats and the user's access rights [5] and the retrieval of the documents of interest [11]. Furthermore Searches based on the actual content are enabled. Classic Information Retrieval (IR) systems, for example, base their searches on comparisons between sequences of characters and they often lead to not accurate and ambiguous results: they not only exclude, from the obtained results, all the documents where the concepts of interest are expressed with terms that are different from the key-word used in the query, but they also present a low level of semantic pertinence with respect to the user information needs, presenting information that doesn't pertain to the user domain of interest.

1.1.1 The Italian Perspective

The strategic plans provided for the actions of e-Government have the aims of establishing cooperation and coordination among the different subjects of Public Administration. In the last decade, Public Administration in Italy has been changing its own organizational structure in order to enable the development of appropriate information systems with respect to the new application requirements,

1993	First definition by law of digital document, art. 491 bis c.p. l. December 23, 1993 n. 547 , Modifications and integration to the norms of the c.p and the c.p.p in topic of computer science crime: "qualunque supporto informatico contenente dati o informazioni aventi efficacia probatoria o programmi specificamente destinati ad elaborarli"
1997	Art. 15, c. 2° l. march 15, 1997, n. 59 (c.d. Legge Bassanini-uno) containing the "Delega al Governo per il conferimento di funzioni e compiti alle regioni ed agli enti locali per la riforma della Pubblica Amministrazione e per la semplificazione amministrativa". This Law has for the first time affirmed the principle of full validity and relevance of computer document, stating that "Gli atti, dati e documenti formati dalla pubblica amministrazione e dai privati con strumenti informatici o telematici, i contratti stipulati nelle medesime forme, nonché la loro archiviazione e trasmissione con strumenti informatici sono validi e rilevanti a tutti gli effetti di legge"
1997	First implemental regulation, d. pres. November 10, 1997, n. 513 . The digital document is defined as a "Rappresentazione informatica di atti, fatti o dati giuridicamente rilevanti"
1999	1999/93/EC Directive on a community framework for digital signatures
2002	Adjustment of the Italian law to EC Directive, D. legisl. January 23, 2002, n. 10
2004	D.p.c.m. January 13, 2004 , "Regole tecniche per la formazione, la trasmissione, la conservazione, la riproduzione e la validazione, anche temporale, dei documenti informatici"
2004	Deliberation CNIPA February 19, 2004, n. 11 , "Regole tecniche per la riproduzione e conservazione di documenti su supporto ottico"
2005	D.legisl. March 7, 2005, n. 82 containing the "Codice dell'amministrazione digitale" (C.A.D.) for the coordination and reorganization of existing provisions about information organizations
2007	D. legisl. April 4, 2006, n. 159 supplementary and corrective rules for C.A.D.

Figure 1.1: The Italian legislative Scenario

by reorganizing itself, implementing its own standards and adopting European and international ones.

Italy, as well as many national systems, greatly needs to arrange appropriate systems able to ensure its growth, development and competitiveness.

There is a real necessity of a de-bureaucratization and simplification of the processes in order to: (i) provide the public and private administrative acts with transparency; (ii) to increase the quality of the offered services; (iii) to decrease the costs of the organization, thus increasing its efficiency.

It is necessary to evolve from systems based on paper documents and manual processes, to information systems focused on processes, which are totally automated and based on digital documents, which are able to optimize and rationalize the use of the human resources involved. A list of most significant regulations about the validity of digital document in bureaucratic contest for dematerialization aim is reported in fig 1.1.

Nowadays, the main instruments achieved, still in evolution, concern electronic signature for the documents legal validity, digital protocol, long term preservation of electronic documents according to regulations, and service of certified electronic mails to give evidence to the posting and receipt of documents.

In Italy, CNIPA (“Centro Nazionale per Informatica nella Pubblica Amministrazione”) has regulated a reference model for the interoperability and the cooperation of the Public Administration named “Architecture of the Public System of Connectivity and Cooperation (PSC)”. Such Model comprehend a set of technological standards and infrastructure services whose objective is enabling the interoperability and the cooperation of the information systems for the fulfillment of administrative actions;

the offered services aim at creating a groundwork to which all the Regions can connect in order to use and distribute services through standard protocols.

1.2 Objectives

In this work we propose a new model of multimedia document, suitable for e-Government activities, that takes into account the requirement of the e-Government applications which, depending on authorities, final users or time, produces different representations of the same multimedia contents. For this reason we model presentation and informative content in a separate way, allowing to solve, among the others things, open problems related to technology evolution, different document format and access rights. The proposed model constitutes the starting point for an information system in the most efficient way, which integrates and processes different multimedia data type (like as images, text, graphic objects, audio, video, composite multimedia, etc.) and, in particular, it allows: *i*) structuring of documents *ii*) automatic information extraction from digital documents; *iii*) semantic retrieval; *vi*) semantic interpretation of the relevant information presented in the document, *v*) storing and *vi*) long term preservation.

The proposed system combines ORDBMS technologies, NLP techniques, proper domain and structural ontologies, and inference rules in order to retrieve the significant concepts related to each documents and to provide querying facilities for users. One important facilities implemented by our system is the possibility to make advanced searches overcoming the barrier imposed by the “keyword-based” traditional query and to allow a “content-based” access to the documental

database, giving great attention to the efficiency aspects, that are strictly related to the usability and the consequent effectiveness of the whole system. The traditional information retrieval systems, based on the comparison of sequences of characters, are in fact able to identify the relevant concepts only if they are expressed within the text with the same terms: the search is always limited to the specific key-words inserted into the query and excludes all the text portions where those keywords do not specifically appear. For instance, if one search for the word “house”, the system will ignore the documents where the words “home” or “residence” appear, even if they represent, in many contexts, the same concept the user is searching for. We exploit semantic characterization of the document content, in order to improve the quality of the information retrieval.

Ontologies play an important role in the process for representation and use of domain specific knowledge[30], by documents metadata annotations for supporting the process of information structuring and retrieval.

The quality of information retrieved is improved by exploiting the possibility to enrich and afterwards refine the list of the retrieved documents by exploiting reasoning techniques on the ontologically-defined relations.

In order to manage the composition of different multimedia data, their semantic relations and the structure imposed for bureaucratic documents, the defined document model is divided in levels, as described in the following:

Data Management Layer: describes the semantic content of each single media element (such as a text fragment or an image), providing functionalities for working on a single media; for example, information extraction and indexing over text is performed in this layer.

Integration layer: describes the relations between the heterogeneous media components of the document, provides functionality for the integration of different data sources. At this layer belonging for example the propriety of a text fragment of referring to an image.

Presentation layer: regulates the way in which the information has to be showed to a single user within a certain context in different times. It provides dif-

ferent representations of the same informative content, according to the formats, the final user's access rights and the technology at disposal.

This approach allows to manage heterogeneous contents, to operate on form and content in an independent way, enabling solutions of open problems related to evolution of the technologies: to give a concrete example, it permits to give an immutable legal validity to the content of a document even if the format of representation changes, evolving with the technology. On different layers of the document, information are semi-automatically tagged according to the concepts contained in the domain ontologies: association among such concepts and their instances, belonging to the document, are picked out. Different ontologies can be used for the tagging process according to the different domains of interest. Besides the Domain Ontology used to formalize the concepts and their relations of interest in the reference domain, it is possible to exploit the defined specialized ontology [18]:

Structure Ontology that describes how information are organized within the document. It models the associations between the internal sections of the document and the set of concepts that can be found in it.

Lexical Ontology that contains the terms of the general language, and can be used to refer wide-ranging concepts presented in the documents, not enclosed in the domain of reference.

Starting from the model, we have proposed an architecture, successively implemented in a prototype system, for the management of the document whole life cycle. It is composed of three main modules: one for the text processing, one for the processing of the other media typologies of data, one for the management of the different formats of presentation, according to the requirements of the E-Government applications. For this dissertation, we have focused on the text processing functionality.

The **Text processing** module aims at extracting the relevant information from text, associating concepts to the terms of the text and defining relations between

them. The text is processed by a series of procedures each of which producing information usable by the next procedures [6]:

Structural Analysis: performs procedures for the text segmentation and the relative classification in order to identify the different sections constituting the structure of the document.

Morpho-Syntactic and Statistical Analysis: performs procedures of language analysis (such as text tokenization and normalization, Part-of-Speech Tagging and lemmatization, complex terms analysis) combined with statistic procedures (such as the computation of opportune indices) enabling the extraction of relevant terms from the corpus to process. These terms and the information about them, refined with the help of domain experts, constitutes a lexicon that are the starting point for the building of the set of concepts that are used for the domain formalization, by means of ontologies.

Semantic Analysis: using the information of the early analyses individuates proprieties and associations among the terms, defining the concepts and the relationships among them, allowing in different phases, ontology building and document annotations.

The **Multimedia Data Processing** module has the aims of classifying each multimedia element, associating concepts from the domain ontology. It is composed of two components implementing innovative methods that have been presented in recent works [19][20]

Analyzer : it identifies the relevant media parts and produces a low-level description that permits to create a series of indices to help the tagging and retrieval procedures.

Classifier : it uses the indexing information to deduce which concepts, from the domain ontology, are to be associated to the media element.

Finally, the **Presentation** module performs the dual task of combining the information about the heterogeneous contents and managing the modalities through

which they are presented to the different users, according to the policy of the Entity (as the Public Administration), the final user's access rights and the technology at disposal. The whole process of document management, performed by the designed architecture, can be divided in three main stages:

Domain formalization: this stage have the aim to codify, with opportune data structures (ontologies) the information of interest pertaining to the domain the documents belong to. Information associated to content are codified in terms of relevant concepts and relations between them.

Document association to the opportune domain of reference: this stage serves to automatically classify the documents given in input, associating to them the domain of pertinence, indicating, thus, the concepts and the relations instantiated within the documents.

Final users utilization: this stage implements the functionalities of document processing offered to the users in order to perform automatic procedures on documents, such as searches by contents, long-term preservation and information representation according to different formats and different access policies.

We have implemented a prototypal version of the system that realizes the described data management procedures, some experimental results are reported which we have carried out for evaluating the impact of the proposed system on the enhancing user effort in automatic information extraction and in juridical documents indexing for retrieval purposes.

1.3 Motivating Examples

To better explain the purpose of the thesis we report and explain tree examples that motivating that can be properly, efficacy and efficiently managed by a system for semantic document processing.

throughout the thesys, the first one will be used as a running example, as discussed in the following.

1.3.1 Juridical Documents

Let us consider the Italian juridic domain, and in particular the notary one: a notary is someone legally empowered to certify the legal validity of a document. Let us suppose to analyze a *buying act*. In real estate market, in Italy and also in some other european countries, when someone has the intention of buying or selling a property, such as houses, pieces of lands and so on, a notary document, certifying the property transaction from an individual to another one, is signed. Such document is generally composed by an *introduction part* containing the caption, a part containing the *biographical data* of the individuals involved in the buying act, a section containing *data about the property* and a sequence containing several rules regulating the sales contract. Consider for example the Italian sales contract fragment, proposed in figure 1.2; an Italian reader can easily detect the areas concerning the caption, the personal data and the property attributes. In a similar way, we propose a system that: i) detects the several sections containing relevant information (segmentation), and ii) transforms the unstructured information within the retrieved section into a structured document, by means of a proper formalization of the information pertaining to it.

1.3.2 Practices of Telemedicine

In the last years Italian P.A. financed several projects aimed at enhancing (and easing) protocols in health boards and medical offices, and at providing continuous and complex health services for patients in critical conditions. These projects falls in the Telemedicine domain, and include, for example, the Unique Centres for Reservation (CUPs - Centri Unici di Prenotazione) of clinical analyses; tele-monitoring systems for high-blood pressure sufferer and for heart patients; or the system for medicine prescription and selling.

All these systems usually require the acquisition of information about all clinical story of patients. For this reason, and to allow for a faster management of users request, a project for providing a single *Digital Case History* for citizens has been financed. This requires a management of clinical data in multiple formats (texts, images, video and audio) and, for older medical reports and clinical

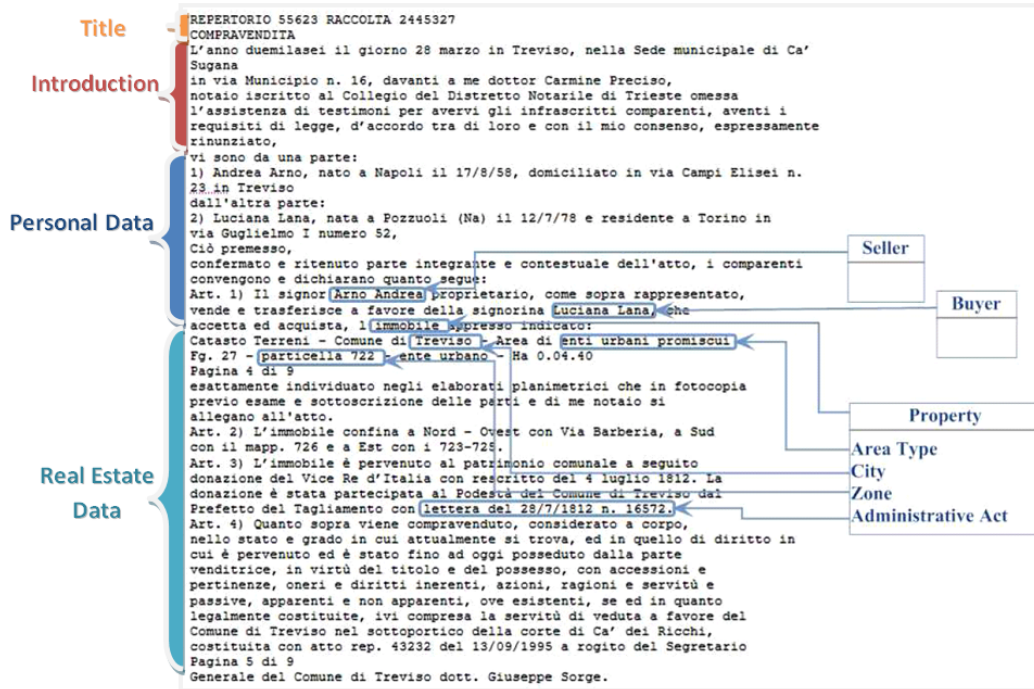


Figure 1.2: An Example of Italian Juridical Document

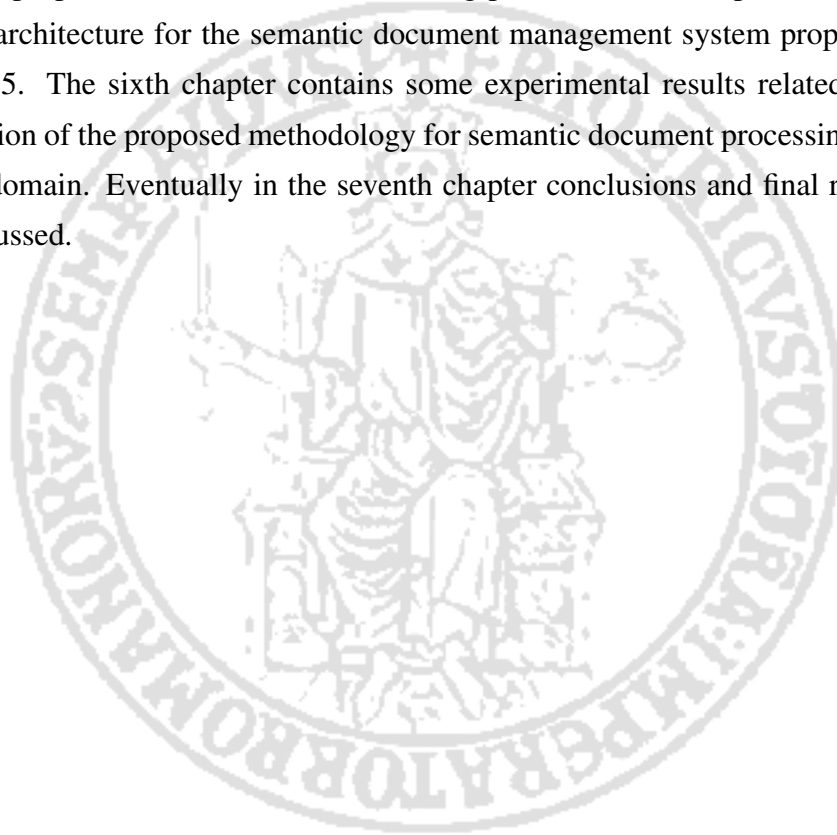
analyses, dematerialization activities have to be enacted for filling digital case histories.

Also in this case, semantic management of data in dematerialization, similarly to previous example, is appealing since it is able to produce structured documents, like is required by some international standards for Digital Case History filling like *Healt Level 7* (HL7).

In fact, like for juridical documents, a Case History is usually composed by sections, where the institute, the medical department and unit, and the doctor and patient names are reported. Then a section which explain actual and previous symptoms, previous recipes, diagnosis and prescriptions are reported. It is simple to notice that in each section different domain terms can be retrieved, like in the prescription section where medical analyses and medical remedies are listed.

1.4 Outline

The thesis is organized as in the following: in the next chapter an overview on Methodologies for Domain Characterization and Formalization is presented. In the third chapter the state of the art related to Document Management System Techniques is reported. In the forth chapter a model for digital document characterization is proposed; such model is the starting point for the description of the the system architecture for the semantic document management system proposed in chapter 5. The sixth chapter contains some experimental results related to the application of the proposed methodology for semantic document processing in juridical domain. Eventually in the seventh chapter conclusions and final remarks are discussed.



Chapter 2

Domain Characterization: Methodologies and State of the Art

2.1 Introduction to Domain Modeling

The design and the development of computer system and software usually imply some descriptions of the reality of interest in the form of conceptual models. Such models have the goal of representing the domain of interest handled by the application, and have particularly been used by programs and databases. Conceptual models intend to describe the relevant aspects of a domain, by mapping the gathered business requirements to the structures of the model and abstracting technical design considerations. Such models are typically used to illustrate the processes, rules, entities, and organizational units that have been identified. Several types of conceptual models are defined on the basis of the purpose of the target application. The most frequently used are:

- Database Models, which are created to describe the conceptual, logic and

physical design or schema of all the information stored in it. Logic model are represented in E-R schema, physical model deals with the conversion of the conceptual model, into schema according to some database language, e.g. SQL or XML.

- Software Applications Model, which are create to model: the functionality of the system from the user's point of view; the structure of the system by using objects, attributes, operations, relationships; and the behavior of the software system. The modeling language is UML, that in the version 2.2, provides fourteen types of diagrams: seven diagram types to model structural aspect whereas the other seven represent general types of behavior, including four that represent different aspects of interactions.

In order to formalize the domain at issue, in different application contexts and within different communities, the computational artifacts are used as conceptual models to capture the knowledge of a particular domain. Information systems, in particular, can use ontologies to get access to such domain knowledge in a computational way. Ontologies have been explored from different points of view, and there exist several definitions of what an ontology is. In the following, the ontology definition, context and use are provided.

2.2 Ontology Definitions

“Ontology” is a word coming from the Greek, formed by *ὄν* *ontos*: of being (neuter participle of *εἶναι*: to be) and *-λογία*, *-logia*: science, study, theory).

It is defined as the philosophical study of the nature of being, existence and reality in general, as well as of the basic categories of being and their relations [?].

The philosopher Plato (427 - 347 BC) was one of the first to explicitly mention the *world of ideas or forms* in contrast to the real or observed objects, which are imperfect realizations (or shadows) of the *ideas*. In “The Sophist” Plato argues that *Being is a Form in which all existent things participate and which they have in common*: the ideas, forms or abstractions are ascribed to the entities which one can talk about, and constitute the foundations for ontology. Some years later, Aristotle, a student of Plato’s, in his “Metaphysics”, outlines the logical background of ontologies, introducing notions such as *category*, *subsumption*, as well as the *superconcept/subconcept* and the consequent concept of *inheritance*. Aristotle can also be regarded as the founder of *taxonomy*, i.e. the science of classifying things, furthermore, he introduced a number of inference rules, called *syllogisms*, such as those used in modern logic-based reasoning systems[2].

In the computer science field, the ontology term does no more referring to the science of the existence, but it refers to the formal specification of a conceptual-

ization, as cited in the various ontology definitions given by Gruber [3].

Definition *Ontology Definition (Gruber 1993)*

An ontology is a formal explicit specification of a shared conceptualization of a domain of interest.

This definition contains two key points:

- the conceptualisation that, being formal, permits some reasoning by computer;
- a practical ontology that is designed for some particular domain of interest.

This definition of ontology is the most cited one but, in literature, other different definitions have been given by different research groups, which often contradict one another.

An ontology aims at providing a formal and explicit description of the concepts in a domain of discourse. The principal constituents of an ontology are concepts, relations and instances: concepts represent the categories and the classes of objects that are relevant in the domain of interest; relations serve to semantically connect concepts and instances; instances represent the named and identifiable concrete objects in the domain of interest, i.e. the particular individuals that are classified by concepts and interrelated by relations.

Many definitions of what an ontology is have been proposed, in particular already in the early years of ontology research, Guarino and Giaretta (1995) raised

concerns that the term “ontology” was used in many acceptions, sometimes even inconsistently. They found at least seven different notions assigned to the term “ontology”: a philosophical discipline, an informal conceptual system, a formal semantic account, a specification of a conceptualization, a representation of a conceptual system via a logical theory, (characterized by specific formal properties or characterized only by its specific purposes), a vocabulary used by a logical theory and, finally, a (meta-level) specification of a logical theory.

They arrived at a definition of ontology weakening the most popular (but sometimes misunderstood) Gruber’s definition.

Definition ((Guarino & Giaretta, 1995) An ontology is a logical theory which gives an explicit, partial account of a conceptualization .

With partial account Guarino means that the formal content of an ontology cannot completely specify the intended meaning of a conceptual element but only approximate it, mostly, by making unwanted interpretations and logical contradictions.

Even if today there is still a lot of inconsistency in the use of the term, in particular at the border between computer science and information system research, we can here report some of the most used ontology definitions.

Definition (Staab and Studer, 2004 [?]) Ontologies consist of concepts (also knowns as classes), relations (properties), instances and axioms.

Definition *Ontology Definition (Staab and Studer)*

An ontology is a 4-tuple $\langle C, R, I, A \rangle$, where C is a set of concepts, R a set of relations, I a set of instances and A a set of axioms.

There exist several approaches of classifying types of ontologies, proposed among others by Lassila & McGuinness in 2001 [(Lassila & McGuinness, 2001)] and by Oberle (Oberle, 2006, pp. 4347).

In this dissertation we characterize the ontology using 3 dimension:

1. Number of conceptual elements in the domain: some domain ontologies are very large so it takes more effort to managed them. But large ontologies can also be unfeasible for use with reasoners that require an in-memory model of the ontology, Often, smaller ontologies are adopted more quickly and gain a greater popularity than large ones (Hepp, 2007).
2. Degree of ambiguity in the conceptualization of the domain: the more a domain is specialized the less ambiguous it is. This means that a specialized concept should be less subjected to misunderstandings since its interpretation is socially shared among the community of experts.
3. Expressiveness of the formalism used to specify the ontology: this can range from a frame-based vocabulary to a richly axiomatized ontology in higher order logic. A higher expressiveness allows more sophisticated reasoning

and excludes more unwanted interpretations, but it also requires much more effort to produce the ontology. Furthermore, it is more difficult for users to understand an expressive ontology, since it requires a better education in logic and more time. Finally, expressiveness increases the computational costs of reasoning.

We report these factors as dimension of a cartesian asses in the fig.2.1[16, 17].

In addition, especially in the context of the Semantic Web, there have been many proposals for an ontology language with a well-defined syntax and formal semantics, such as OIL [Horrocks et al., 2000], RDFS [Brickley and Guha, 2002] or OWL [Bechhofer et al., 2004].

“”

2.3 Problems of interpretation in non-specialized domains: the ambiguity of natural language

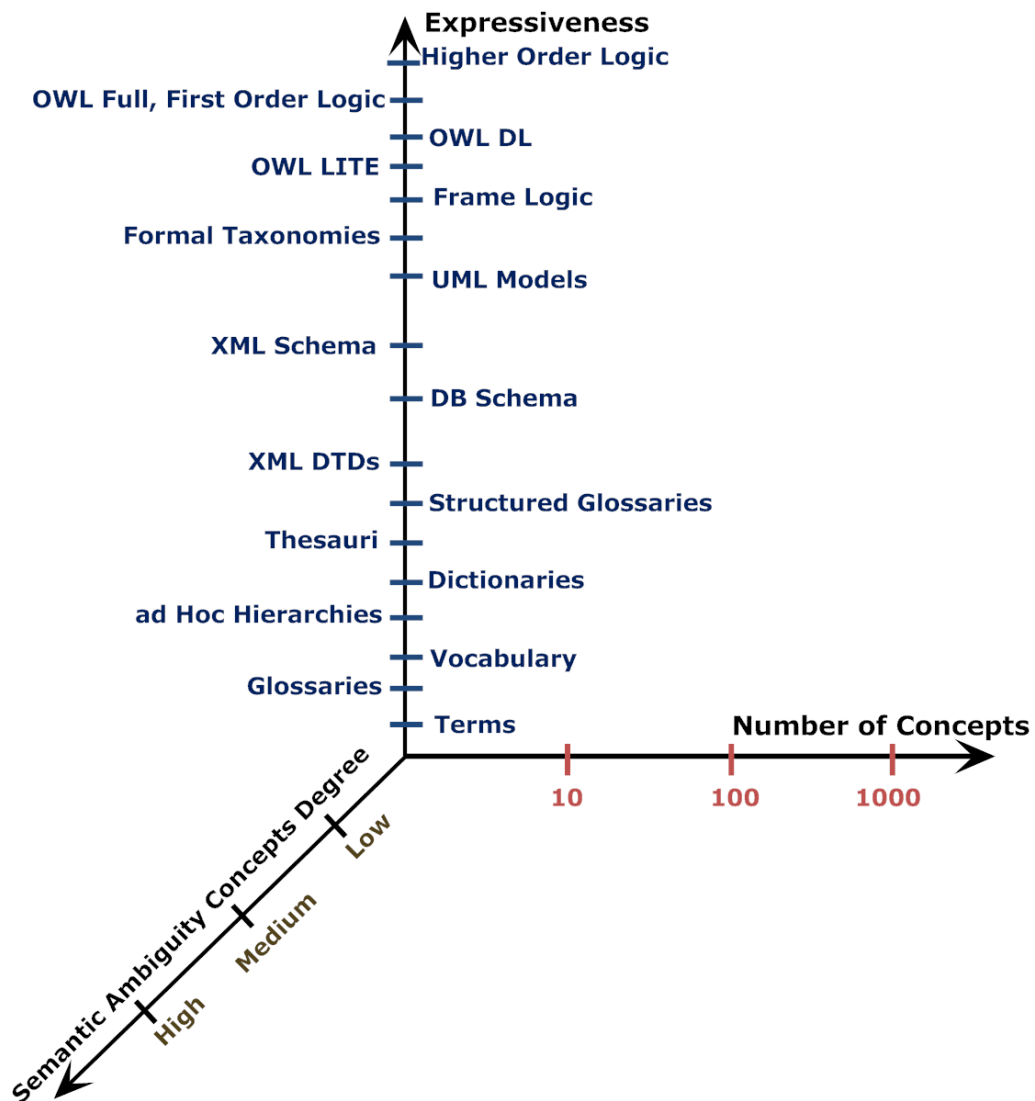


Figure 2.1: Domain Knowledge Formalization

Il	successo	fa	bene
Article	Verb: Past-Part	Verb	Adjective
	Common Noun	Noun	Noun
			Adverb

Table 2.1: Example of word category ambiguity for the Italian language

Several problems prevent documents in natural language to be comprehended through automatic procedures, in particular problems due to the ambiguity and the indefiniteness which make expressions compatible with various interpretations. Ambiguity can affect all the levels of the language, in particular the morpho-syntactic, syntactic and semantic ones.

At a first level, there could be problems affecting part-of-speech tagging: given a sequence of words, each word can be tagged with different categories (Tamburini, 2000).

In the example above, the disambiguation of a lexical item is enabled by the linguistic context (for example, the word “successo” is disambiguated as common noun since preceded by an article), by taking into account the POS category of the preceding or following words. However, it is also to take into account that even the preceding word can be ambiguous or that the disambiguation of a form can require further semantic or pragmatic knowledge.

Automatic POS tagging is a general problem of *word-category disambiguation* involving two kinds of difficulties: i) finding the POS tag or all the possible

tags for each lexical item; ii) choosing, among all the possible tags, the correct one. The first problem can be solved by using a glossary or a lexical list as reference; the second one, instead, can be solved by using: i) contextual evidences, that is examining the context where the word is used (linguistic approach); ii) probabilistic evidences starting from a tagged corpus to be used to train a tagger (statistical approach). Many researches have been conducted on the problem of automatic pos tagging and different have been the approaches used (linguistic, statistical and hybrid) and the models implemented. Among the principal techniques are: stochastic models (Charniak et al. 1993; Carlberger, Kann 1999, Cutting et al. 1992; Dermatas, Kokkinakis 1995; Deroose 1988; Kupiec 1992), rule-based models (Greene, Rubin 1971, Voutilainen 1995), hybrid systems (Brill 1992, 1994, 1995), memory-based models (Daelemans, Zavrel 1996), decision trees (Màrquez, Rodríguez 1997a,1997b; Schmid 1994). Brill e Wu (1998) combine the output of different taggers to obtain the best performance by means of a vote mechanism: for each word is selected the tag that has been chosen by the higher number of taggers (majority voting). Among the works developed specifically for the Italian language are De Mauro et al. (1993), for stochastic taggers, and Delmonte et al. (1997) for rule-based taggers.

At a syntactic level, there are problems affecting the disambiguation of syntactic structures: it is to note that some sentences are, in fact, susceptible of different interpretations, that's why they can be associated to different *parse trees*. This

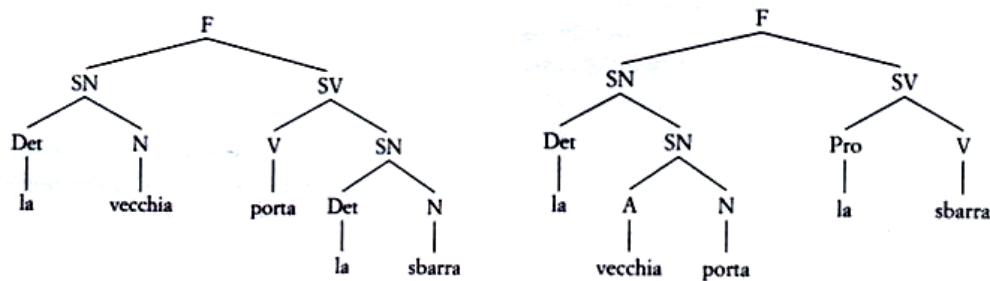


Figure 2.2: Example of syntactic ambiguity

is the case, for instance, of an Italian sentence like “La vecchia porta la sbarra” to which two parse trees can be associated (Fig. 2.2) since interpretable in two ways: i) an old woman brings a bar (1st parse tree); ii) an old door blocks something (2nd parse tree). In the figure: F (ITA: frase) corresponds to the English S – sentence –; SN (ITA: sintagma nominale) corresponds to the English NP – noun phrase –; SV (ITA: sintagma verbale) corresponds to the English VP – verb phrase –; Det, N and Pro stand respectively for determinative, noun and pronoun.

At a semantic level we have firstly to consider the *unpredictability* with which the word meanings develop and get organized. Meanings are internally organized in senses and very often the senses of a same word get specialized in very different and unpredictable ways. Another aspect related to the organization of meanings is their *extensibility*, that is the capacity to develop for a same word new senses to its meaning in order to meet specific communicative requirements. Secondly, the presence of homonyms and polysemous words is another aspect representing a problem for interpretation in a computational field.

If for a human interpreter these characteristics are normal and easily to manage, for a computer the matter is different since the management of these issues require a great quantity of elaboration to implement operation of disambiguation. Many algorithms of word-sense disambiguation (WSD) are dictionary and knowledge-based. These algorithms operates by means of explicit knowledge-bases since they use resources contained within machine readable dictionaries, thesauri, computational lexicons, ontologies. Algorithms of Gloss Overlap (Lesk, 1986; Banerjee and Pedersen, 2002) belong to this approach: they base on the hypothesis that there is some kind of relation between the words that are used together within a sentence. This relation can be determined by observing for each word of the sentence all the possible definitions in a dictionary: a word is correctly disambiguated by comparing all its definitions with the definitions of the other words in the sentence and choosing the one having the higher lexical overlap. Supervised algorithms of WSD, instead, require no access to explicit knowledge since they operate by means of statistical criteria taking into account the linguistic context of words obtained from training corpora. They base on the thesis that the local context can provide evidences for the sense disambiguation: these evidences are obtained from hand-tagged corpora, already containing information about the sense of the words and their relations. Among supervised algorithms of WSD there is the Most Frequent Sense (MFS) algorithm which disambiguates a word by assigning to it the most frequent sense that has been computed within the training

corpus. Finally, there are unsupervised algorithms of WSD that find the correct sense of a word by computing a similarity measure between the target word and the other words within its local context. They base on the thesis that similar senses occur in similar contexts. In this case the sense of a word can be obtained from the text by clustering the occurrences of the word by means of these similarity measures. This process creates lexical chains, that are chains of words semantically linked by means of a relation of cohesion. Each occurrence of the word must belong to one and only one chain. Algorithms belonging to this approach are Morris and Hirst's algorithm (1991) that use a thesaurus as knowledge-base to extract the relations between terms, and Hirst and Stonge's algorithm (1998), that use WordNet as source for relations.

A lexical expression can therefore contain a certain amount of ambiguity, which enables two or more attributions to it: what, in any form, represents aspects of the language incalculability is, thus, managed with great difficulty by a machine.

Generally, problems for an automatic document processing come from the strong interaction and inter-dependence among the syntactic, semantic and pragmatic levels, which make flexible and dynamic the use of the language signs: word senses have uncertain boundaries and very often they change according to the interactions built with other elements within the contexts where they can occur and according to the extra-textual context. Therefore, to describe a document and to

understand its contents it is necessary to identify not only the single signs but even the relations these signs keep up between them, firstly at a syntactic and semantic level and, secondly, at a pragmatic level, that is to say the relations the signs have with the external context and in general with the domain the document pertains to. The semantic dimension, indeed, permits to consider as acceptable only some of the possible syntactic interpretations, and the pragmatic dimension permits to solve many semantic indefiniteness.

Ambiguity can then be solved by resorting to the knowledge of both the co-text and the domain of reference where the texts is placed and used: in this sense, the domain becomes a real encyclopedia functional to the interpretation of the document sense. Not only it enables an immediate interpretation of the language signs but, considering their possible implications, it also permits further interpretations: each expression can, in fact, be subjected to a semantic interpretation and each interpretation can open to other meanings. The encyclopedic knowledge, then, provides instructions to interpret in the most complete way the document sense.

This is important above all when dealing with specialized domains which produce their own documents in their own language variety (or sublanguage): in such domains (like the bureaucratic one) the interpretation of document data is generally unique, given the technicalities introduced in the sublanguage which reduce the problems due to ambiguity and incomprehension.

2.4 Specialist languages and specialized domains

A specialist language represents a sub-variety of the common and general language: it adds to the basic data of the general language specialist data, in relation to the specificity of the concepts dealt, in order to provide the experts of the domain with a technical and rigorous terminology, so to ensure a communication without ambiguity.

Rigour and clarity represent the important characteristics of a specialist language: the former is functional to the possibility of determining the document contents in a univocal way; the latter is functional to the possibility, for the receivers, of an easy access to these contents. Consequently, rigour and clarity depends on the terminology used in the domain: a technical word (or term) must determine its sense in the most rigorous way and convey one single meaning.

Specialist languages aim at an ideal of monosemy, that is a univocal relation between concept expressed and term designating it: each designation must exclusively represent the concept at issue.

Therefore, these languages need to create their own terminology, that is to say their own set of specialist (technical) words.

A term, or terminological unit, is the designation of a concept in a specialist language. This designation can be:

1. a word belonging to the common language that has been assigned with a

new and specialist meaning (*redefinition or technicalized word*): this is an exemplification of a specialist re-use or even sense extension;

2. a word that exclusively belongs to the specific domain (*technicality*): it has a univocal meaning and doesn't occur outside the domain.

Structurally, a term can correspond to:

1. a simple term corresponding to a single word (even if derived or composite), delimited from the other words by two blank spaces;
2. a complex term, composed of two or more words separated by blank spaces forming an expression conveying a complete and autonomous sense.
3. an acronym, an abbreviation, a formula.

Sublanguages, then, can produce new words and expressions or assign a new and a more specialized sense to words already existing in the standard language. Operations of redefinition and technicalization, therefore, produce neologisms of sense which serve to reduce the risks coming from bad interpretations. Operations of derivation, composition and abbreviation, as well as lexicalization, can create, instead, neologisms of form that even serve to characterize the specialist language.

A neologism can become a specialist term of a domain only if it conveys the content of the expressed concept.

Within a specific domain, therefore, a term presents peculiar characteristics:

1. it is univocally related to a specific concept of the domain;
2. it is regularly used to designate a specific concept within the documents pertaining to the domain

Within a specific domain, a specialist concept can be recognized by means of:

1. the set of characteristics describing it in any corpus pertaining to the domain itself;
2. a definition distinguishing it from other concepts;
3. a regular association with a designation.

On its side, a term is recognized by means of a regular association with a set of characteristics able to define the concept it designates. There is, therefore, a semantic stability linking the concept to the term.

Complex expressions are very frequent within specialized domains, given the specificity of the matters to deal: generally they correspond to phrase structures and are the output of technical uses; in particular, they often represent specialized designations of more general concepts.

Therefore, syntagmatic relations are evidence, at a deeper level, of sense relations: words can regularly co-occur because of their intrinsic sense which make them conceptually associated (isotopy).

It is therefore important, while analyzing a specialist text, not to lose the overall sense of these syntagmatic sequences dispersing the single lexical items: it is necessary to process the complex term as autonomous unit of analysis. The identification of these sequences of words is then fundamental for the comprehension of the text: they obviously depend on the semantic of the text and catching them automatically is far from being simple.

Their recognition relies principally on human intervention and involves two principal steps: i) the identification of phrase structures; ii) the selection of the relevant structures designating meaningful concepts of the domain. Semi-automatic techniques in this sense are the key-word-in-context analysis, the co-occurrence analysis and the analysis of the repeated segments (Bolasco, 1999, 2004).

A central aspect for a correct document interpretation is, then, the continuous resorting to the linguistic and extra-linguistic knowledge: all texts are riddled with more or less shared knowledge, some of them are general and common, others depends on our encyclopedia, which works as a hypothesis regulating the interpretation according to the domain of use.

Thus, it is possible to state that the comprehension of specialist documents causes: i) less problems than the comprehension of more general texts since, being more rigorous, they reduce semantic ambiguity; ii) more problems of comprehension for people who are not expert of the domain.

2.5 Concepts and Relations detection

In the field of semantic processing of documents, strategies for *text analysis* and *extraction of knowledge*, in terms of relevant information, are required in order to provide a terminological and conceptual representation of documents.

Knowledge extraction from texts is a fundamental task in the semantic processing scenario but it is also difficult because it is strongly connected to:

1. the personal way by which document authors have made knowledge explicit or implicit within text;
2. the amount of knowledge a reader requires to interpret text contents.

Text data are analyzed for comprehension and transformed into information, that is to say data is transformed into relevant concepts with respect to the particular domain of interest. Concepts identification firstly requires the ability to identify, within the text structure, the entities the refer to, and in second place the ability to identify properties characterizing concepts and relations among them (Dell'Orletta *et al*, 2008\cite {DellOrletta2008}).

The automatic identification of concepts from text data involves several morpho-syntactic problems. Problems are also related to semantic ambiguity, that generally derives from the dynamism and the flexibility of the language signs uses.

A text is the product of a communicative act resulting from a process of col-

laboration between an author and a reader: authors use language signs in order to codify meanings; readers decodes signs and interprets their meanings by exploiting knowledge of:

1. the extra-textual context and, more in general, his *encyclopedic knowledge* involving the domain of interest;
2. the infra-textual context, which consists in relations at a morphologic, syntactic and semantic level.

Thus, the activity of knowledge extraction from texts comprehends different kinds of text analysis methodologies, aiming at recreating the model of the domain texts pertain to.

The state of the art in this field is related to techniques of Natural Language Processing (NLP) and to cross-disciplinary perspectives including Statistical Linguistics[Butler C.S. (1985); De Mauro T. (1961); Rizzi A. (1992)] and Computational Linguistics[Biber D. et alii (1998); Habert B. at alii (1997); Kennedy G. (1998); Spina S. (2001)], whose objective is the study and the analysis of natural language and its functioning through computational tools and models. In particular, for the analysis of limited textual universes, as well as sectorial areas, specific disciplines have been developed, like Corpora Linguistics[Biber D. et alii (1998); Habert B. at alii (1997); Kennedy G. (1998); Spina S. (2001)] and Textual and Lexical Statistics[Bolasco S. (1999); Bolasco S. (2004); De Mauro T. (1980, 1997); La Torre

M. (2005); Lebart L. et alii (1998); Muller Ch. (1991)].

2.5.1 Relevant Terms Recognition

Term-extraction is the first operational stage in the activities of automatic document processing and derivation of knowledge from texts. It is focused on the analysis of the lexical items since they hold specific conceptual meanings. Words are used to identify fundamental concepts of a specific knowledge domain: they have their realization within texts and their relations constitute the semantic frame both for documents and for the domain itself.

The main goal of this stage is to find relevant and peculiar terms in order to define a terminological peculiar lexicon for documents collections.

In this phase we pay particular attention to the analysis not only of simple words but also of complex words, which are syntagmatic combinations of terms. The analysis leads to identification of specific domain concepts within documents.

Methods for term extraction from texts can be divided in three main categories:

Linguistic, Statistical and Hybrid methods.

Linguistic methods exploit linguistic knowledge about term formation in order to find terms in a text. They are generally language-dependent. These methods are based on heuristic rules, and help the following activities:

1. tokenization and normalization, in order to identify tokens and harmonize spelling and capitalization;

2. part-of-speech tagging, in order to filter terms for extracting only the categories of interest, such as nouns and verbs;
3. word-stemming, in order to convert words to their root form;
4. lemmatization, in order to restore words to a dictionary form;
5. identification of phrase-structures that can represent specialization of more general concepts, such as , for example the Italian expression “imposta da bollo” (duty stamp in english).

Statistical methods are the base for the analysis of word occurrences within texts. They measure the weight of a candidate term. 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 other ones. Two main characteristics determinate by this methods are: *termhood* and *unithood*.

Termhood measures the degree by which a term is related to the specific concept of the domain and it is based on the frequency of occurrences (Kageura et alii, 1996).

Unithood is useful to detect complex terms forming a unique segment, measuring the significance of the words occurring together. Standard statistical techniques are mutual information and log-likelihood (Ziqi Zhang et alii, 2008; Daille et alii, 1994).

Hybrid methods use a linguistic filter, based on part-of-speech tags, to extract a set of candidate terms. Then statistical methods are used to assign a value to each candidate term.

Pure statistical approach, in fact, produces high values of semantic precision with respect to the corpus contents but poor values of word recall with respect to the language of the domain (Lame).

In order to extract the peculiar words from a document collection with respect to the specific domain of interest, these methods provide a comparison with lexical external resources, such as glossaries and lexicons, and they are usually divided into several steps, which are described in the following.

Text Preprocessing

The main goal of this stage is the extraction of relevant units of lexical elements to be processed in following phases.

Text tokenization

Tokenization consists in the segmentation of the text into minimal units of analysis, defined *tokens*, that, according to each particular case, can correspond to simple or complex lexical items, including compounds, abbreviations, acronyms and alphanumeric expressions.

Text tokenization, then, requires, several sub-steps:

1. graphemic analysis, used for defining the set of alphabetical and non-alphabetical signs actually used within the text collection in order to verify the presence of possible mistakes when sign are not planned in the allowed language;
2. disambiguation of punctuation marks, which can be usually considered independent tokens (for example in the case of end of sentences) or not (for example in acronyms and abbreviations);
3. separation of continuous strings (i.e. strings that are not separated by blank spaces), which have to be considered independent tokens: for example, in the Italian string “c’era” there are two independent tokens (c’ + era);
4. identification of separated strings (i.e. strings that are separated by blank spaces) which have to be considered complex tokens and therefore single units of analysis: Examples are proper names (like “Mario Rossi”, “Reggio Calabria”,etc.), monetary expressions (like “3 euros”), measures (like “23 kg”), dates (like “1° Gennaio 1948”), addresses (like “via Madonnelle 16”), laws (like “dpr 28 dicembre 2000, n. 45”), etc.

This segmentation can be performed by means of special tools, called *tokenizers*. They are composed of two fundamental components:

1. glossaries listing well-known expressions to consider as tokens;

2. mini-grammars containing heuristic rules (in the form of regular rules), which are manually written by experts.

The combined use of glossaries and mini-grammars ensures high levels of accuracy. However results depend on the kind of text and language used: texts which are full of acronyms or abbreviations can increase the percentage of mistakes. Consequently, the glossary and the mini-grammar should be adapted to the characteristics of the issued domain.

Text Normalization

Generally lexical expressions that have to be considered equivalent, can be found within the same document in different forms. This is the case, for example, of identical words written in small and capital letters, compounds and prefixed words that can be (or not be) separated by hyphens, dates that can be written in different ways (“1 Gennaio 1948” or “01/01/48”), acronyms and abbreviations (“USA” or “U.S.A.”, “pag” or “pg”), etc.

Normalization involves a series of problems. For example, the transformation of capital letters into small letters makes the identification of the beginning of a sentence difficult. The same is for the distinction between a proper name of person (like the Italian “Rosa”) and a common noun of a flower (like the Italian “rosa”). Another example is the distinction between an acronym (e.g. “USA”) and a verb (e.g. “usa”, 3rd sing. pers. of the Italian infinitive “usare”).

Normalization can be automatically performed by:

1. comparing the document collection to external lexical lists, for the recognition and the standardization of particular expressions (like well-known abbreviations and acronyms, toponyms, as well as grammatical phrases and specific noun phrases);
2. setting proper parameters in order to uniform the different forms. An example is the reduction of capital letters into small letters according to some pre-arranged conditions, when the noun is located after some punctuation marks when it starts a new paragraph.

Morpho-syntactic analysis

The main goal of this stage is the detection of the category whom the words, both simple and complex forms, belonging to; in order to reduce the list of candidate terms on the sole category of interest, with the aim to extract of the only relevant information

Part-of-speech tagging

Part of Speech (POS) Tagging is a basic and a well-known problem in Natural Language Processing: it consists in the assignment of a grammatical category (noun, verb, adjective, adverb, etc.) to each lexical unit identified within the text collection.

The classic morphology of the Italian language identifies nine parts of speech:

1. five *variable parts of speech* (since susceptible to inflection): noun, verb, adjective, pronoun, article;
2. four *invariable parts of speech*: adverb, preposition, conjunction, interjection.

Beyond this “structural” distinction, there is another more “semantic” distinction:

content words (nouns, verbs, adjectives and adverbs): this is an open and productive class of words, which can be enriched with other lexical items. Generally, nouns are indicators of people, things and places; verbs serve to denote actions, states, conditions and processes; adjectives are indicators of properties or qualities of the noun they refer to; adverbs, instead, represent modifiers of other classes (place, time, manner, etc.).

grammatical (functional) words (articles, prepositions, conjunctions): this is a closed and static class of words, generally frequent in language use.

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 (Tamburini, 2000).

In the example above, the disambiguation of a lexical item is enabled by the linguistic context (for example, the word *success* (in Italian “*successo*”) is disambiguated as common noun since preceded by an article), by taking into account the POS category of the preceding or following words. However, it is also possible that even the preceding word can be ambiguous or that the disambiguation of a form can require further semantic or pragmatic knowledge.

Automatic POS tagging is a general problem of *word-category disambiguation* involving two kinds of difficulties: (i) finding the POS tag or all the possible tags for each lexical item; (ii) choosing, among all the possible tags, the correct one. The first problem can be solved by using a glossary or a lexical list as reference, which gives all the terms and the respective tags that can be associated to them; the second one, instead, can be solved by using: (j) contextual evidences, that is examining the context where the word is used (linguistic approach); (jj) probabilistic evidences starting from a tagged corpus to be used to train a tagger (statistical approach)

Many researches have been conducted on the problem of automatic pos tagging and different have been the approaches used (linguistic, statistical and hybrid) and the models implemented. Among the principal techniques are: stochastic models (Charniak et al. 1993; Carlberger, Kann 1999, Cutting et al. 1992; Dermatas, Kokkinakis 1995; Deroose 1988; Kupiec 1992), rule-based models (Greene, Rubin 1971, Voutilainen 1995), hybrid systems (Brill 1992, 1994, 1995), memory-

based models (Daelemans, Zavrel 1996), decision trees (Màrquez, Rodríguez 1997a,1997b; Schmid 1994). Brill e Wu (1998) combine the output of different taggers to obtain the best performance by means of a vote mechanism: for each word is selected the tag that has been chosen by the higher number of taggers (majority voting). Among the works developed specifically for the Italian language are De Mauro et al. (1993), for stochastic taggers, and Delmonte et al. (1997) for rule-based taggers..

POS tagging is performed by comparing the vocabulary of the document collection with an external lexical resource, whereas the procedure of disambiguation is carried out through the analysis of the words in their contexts of occurrence.

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 textual element, it is possible to locate its occurrences in the text and, so, the textual parts preceding and following each one of its occurrences: in particular, at a lexical level, the co-text of a word X coincides with a certain number of preceding and following words, which constitute its left and right neighbourhood.

This kind of analysis, then, permits to visualize the use of the words in their contexts of occurrence in order to disambiguate their grammar category. In this way, the ambiguity between noun and adjective in the Italian word “pubblico” can be solved by observing the categories of the preceding or following words. In the

case in point, the presence of an article, a preposition or a noun: in the first two cases the word at issue is a noun, in the last one it is an adjective.

The ambiguous form is then firstly associated the set of possible POS tags, and then disambiguate by resorting to the KWIC analysis. Here the set of rules defining the possible combinations of sequences of tags, proper of the language, enables the individuation of the correct word category.

Further morphological specifications, such as inflectional information (as gender, masculine/feminine, and number, singular/plural), are then associated to each word.

Finally text lemmatization is produced 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.

Analysis of syntagmatic combinations of words: identification of phrase structures and lexicalization

Our approach is based on the idea that words and their *syntagmatic* combinations convey the conceptual contents of a text and, more in general, of the respective ontological domain, consequently the analysis of the syntactic combinations of words is a fundamental prerequisite.

It is very common to find, within a text, words occurring regularly together in the form of lexical segments and often producing complex words to be considered

as single units of analysis. These complex words coincide, at a syntactic level, to *phrase structures* that often correspond to Italian technical expressions, outcomes of standardized language uses, as in jargons.

On the base of their semantic relevance, two kinds of complex forms can be considered:

1. **content** lexical segments:

- (a) technical terms, often coinciding with compounds and noun phrases, such as the Italian forms “consiglio d’amministrazione”, “collegio notarile”, “base imponibile”
- (b) phrasal verbs, such as the Italian forms “avere ad oggetto”, “fare eccezione”, “fare riferimento”;
- (c) idiomatic expressions, such as “ad hoc”;

2. **grammatical** phrases with function of:

- (a) adverb, such as the Italian expressions “di nuovo”, “in realtà”, “più o meno”;
- (b) preposition, such as the Italian expressions “a margine di”, “a carico di”;
- (c) conjunction, such as the Italian expressions “il fatto che”, “dal momento che”.

The recognition of these syntactic combinations involves two principal steps: (i) the identification of semantically cohesive segments; (ii) the selection of the relevant segments with respect to the domain.

The core hypothesis is that if two or more words form a complex term within a certain domain, it is very probable that in that domain they tend to occur together. This probability is functional to the co-occurrence of the words themselves: if a pair of words occur in the text more often than one would expect, then their co-occurrence can be considered as statistically significant.

Semi-automatic techniques in this sense are the key-word-in-context analysis, the co-occurrence analysis and the analysis of repeating segments (Bolasco, 1999, 2004).

The first one links the description of the corpus vocabulary to the concrete use of the terms in the co-text.

The second one points out the principal associations between the words counting how many times two forms are close together. The computed value constitutes the co-occurrence (or co-frequency) between the two forms. The analysis involves the selection of some parameters, such as:

1. a minimal threshold of occurrence or a list of words to consider in order to determine the vocabulary subset on which to perform the search of co-occurrences;

2. the extent of the neighbourhood, that is the number of words inside which it is possible to notice a co-occurrence.

This kind of analysis can be useful to identify the valence of a verb, or for example, the nouns taken by a verb within the same text.

The analysis of repeating segments is based on the selection of several parameters, as the:

1. marks for delimiting the textual portions where the segments are to be extracted;
2. minimum frequency threshold of the words belonging to the segment in order to determine the vocabulary subset on which to perform the search of the segments;
3. maximum number of words within the segment in order to determine its length;
4. minimum frequency threshold of the segment (that obviously should not be lower than the minimal frequency threshold of the words within the segment itself);
5. the skimming of the redundant list of segments obtained by means of computation of a measure of association among the words composing the repeating segments.

A method to compute this association rate on the base of the tendency of words to co-occur in a text is the *Index of Significance (IS)* (Bolasco, 1999, 2004) which permits to filter the list of redundant segments in order to extract the only relevant and meaningful sequences, in accordance with their capacity of absorption of the occurrences of the compositional words. A word is said completely absorbed by the segment, if all its occurrences appear within the segment; if the most part of the word occurrences appear outside the segment, then the word is not to consider useful to produce a segment (it is “little absorbed”): the higher is the segment power of absorption of the single lexical components, the more a segment is relevant. The IS index adds the ratios between the frequency of the segment and the frequency of the L words belonging to the segment, comparing then the sum obtained to the number P of “*content words*” (not “*grammatical words*”).

$$IS = \left[\sum_{i=1}^L \frac{f_{segm}}{f_{g_i}} \right] \cdot P$$

The IS index is strongly conditioned by the number of content words composing it, therefore it highlight the longer segments, which are not necessarily the more frequent. To obviate this problem, there is the *Relative Index of Significance – ISR* – whose value, which is obtained by dividing the IS index by its maxim value (P^2) of the number P of “*content words*”, oscillates between 0 and 1:

$$ISR = \frac{IS}{P^2}$$

Recurring to ISR indexes, integrating with the human intervention, enable the

identification of a list of relevant complex items.

This list can be further extended by including complex terms of higher order: the procedure of extraction of complex terms can be, in fact, iteratively applied re-projecting onto the segmented text the complex terms previously extracted. For example, if during a first stage the complex term “Economic Community” has been extracted, then a new complex term can be extracted “European Economic Community”, which includes the term previously acquired.

A process of **lexicalization** is then performed on the list of relevant segments in order to turn them into single compact lexical unit, that is single tokens.

2.5.2 Extraction of the relevant terminology

This stage proposes to identify from the list of candidate terms previously obtained, those lexical expressions conveying effective relevant concepts for the document collection at issue and, in general, for the domain of reference.

More in details, Two different kinds of resources can be used:

1. *endogenous resources* (corpus-based), for the creation of a word-set containing the statistically significant and corpus representative key-words;
2. *exogenous resources* (non corpus-based), for the extraction of a word-set containing the terms which are typical of the domain at issue because representing specific domain referential entities.

In both cases, the attention is focused on the word categories of interest, such as nouns, verbs and adjectives, as well as relevant phrase structures.

Our approach, illustrated in the Chapter 5 consist in the integration of the two strategies, evaluating the representation degree of the selected terms in respect of the corpus, in order to filter and specialize the results.

Endogenous resources: the computation of the TFIDF index

The main goal of Information Retrieval techniques is the extraction of relevant information from documents collections.

In order to ensure a good correspondence between query searches and results, the identification of characterizing key-words is required.

As a matter of fact, not all words are equally useful to describe documents: some words are semantically more relevant than others. In endogenous approach the semantic relevance is caught by the assignment of TF-IDF index (*Term Frequency - Inverse Document Frequency*), computed on the corpus vocabulary and on the base of the term frequency and the term distribution within the corpus. TFIDF index, in fact, takes into account:

term frequency (*tf*), corresponding to the number of times a term occurs in the collection: the more a term occurs in the same document, 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 (Dagan et alii, 1994).

inverse document frequency (*idf*), concerns the term distribution on the corpus, on the basis of the term frequency (*tf*) and the term distribution within the corpus (*idf*). 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 documents contain that given term, the less discriminating it is. This index is often used as a baseline (Ziqi Zhang et alii, 2008) or as one of several features to determine the termhood (Medelyan et alii, 2006).

Therefore, TFIDF enables the extraction of the most discriminating lexical items because frequent and concentrated on few documents. This statement is summarized in the following ratio:

$$Wtd = ftd * \log N/Dt$$

where Wtd is the evaluated weight of term t in document d ; ftd is the frequency of term t in document d ; N is the total number of occurrences within the corpus; Dt is the number of documents containing the term t .

However, there's to say that, in our running example, legal terms may present high or low rate of TFIDF, that's why a pure statistical approach is useful to extract *statistically significant* words whose semantic specificity and peculiarity is

evaluated with regard to the topics dealt in the corpus. Statistical indexes, in fact, produce high rates of semantic precision with respect to the corpus contents but poor rates of lexical recall with respect to the domain language: statistical indexes are useful to identify index terms, but they are not so effective for distinguishing domain terms from non-domain terms.

Exogenous resources: lexical comparisons

When dealing with specialized domains documents are expressed by using several language varieties (or sublanguages): these are specialization of standard languages since they add specialized data to the basic ones present in the language, in relation to the specificity of the concepts dealt. This allows for providing a technical and rigorous vocabulary to domain experts.

Each sub-language needs its own vocabulary. It can be defined by introducing new words and expressions, or by assigning a new or more specialized sense to words already existing in the standard language. A specialized variety, in fact, is characterized by the presence of technicalities and technicalized (redefinitions) expressions.

Technicalities are words that exclusively belong to the specific domain: they have a unique meaning and don't occur outside the domain.

Redefinitions, instead, are words that belong to the common language but they are assigned with a new and specialized meaning within the domain: they are

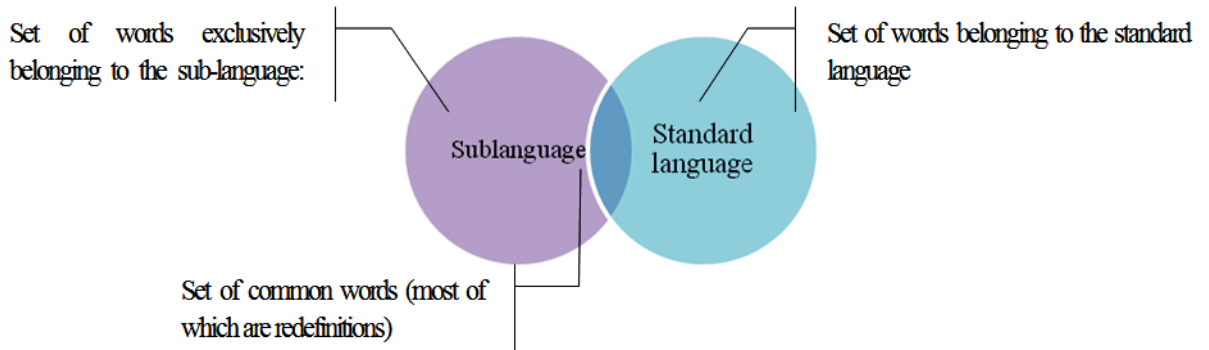


Figure 2.3: Intersections among language varieties

exemplifications of a specialized re-use. A sublanguage can then be enriched with new expressions or can adapt new senses and contexts to expressions already existing. These operations allow for reducing ambiguity in words interpretations.

A text is the output of a language system: the vocabulary of a specialized domain is made of a sub-set of the standard vocabulary. Therefore, the comparisons between different vocabularies would produce interesting results on the kind of words used within the texts.

The comparison with one or more external (general or specialized) lexical resources of reference, usually built with the help of the domain experts, represents a different approach from statistical one, but it can be integrated with it in order to filter the results previously obtained.

Starting from the list containing the morpho-syntactic categories of interest, or from the results obtained by computing the TFIDF index, it is possible to proceed with a comparison with an external lexical resource in order to obtain, for

example, the common terminology, the original terminology of one of them or the union of the terms.

For instance, the comparison between the lemmas extracted in the vocabulary of a legal text corpus and the lemmas of a legal dictionary would produce different outputs. In particular the difference relies in the list of common words, which is the set of terms surely pertaining to the legal domain and leading to fundamental domain concepts.

A comparison with a general dictionary, instead, would produce, among the others, a list of common words, useful for identifying the set of redefinitions.

It is also possible to evaluate the differences in the words occurrences, by comparing the relative frequency of word occurrences in the list of reference.

All this lead to the evaluation of a *standardized difference* index indicating the measure of the word over or under represented: the higher is the value of this difference, the more typical and peculiar is the word with respect to the text at issue. The computation of this value, then, permits the identification of forms that are significantly represented within the text.

This measure is expressed by the following ratio z_i :

$$z_i = \frac{f_i - f_{i*}}{\sqrt{f_{i*}}}$$

where f_i is the number of standardized frequencies of the word i and f_{i*} is the

correspondent value in the list of reference.

This index, then, computes the term peculiarity in terms of positive (over-representation) and negative (under-representation) specificity: the first index is connected to the more frequent words and identifies the peculiar forms, the latter is connected to the less frequent (or even rare) words.

The whole process of the extraction of relevant terminology, that exploits the presented linguistic and statistical analysis, is illustrated in fig. 2.5.2.

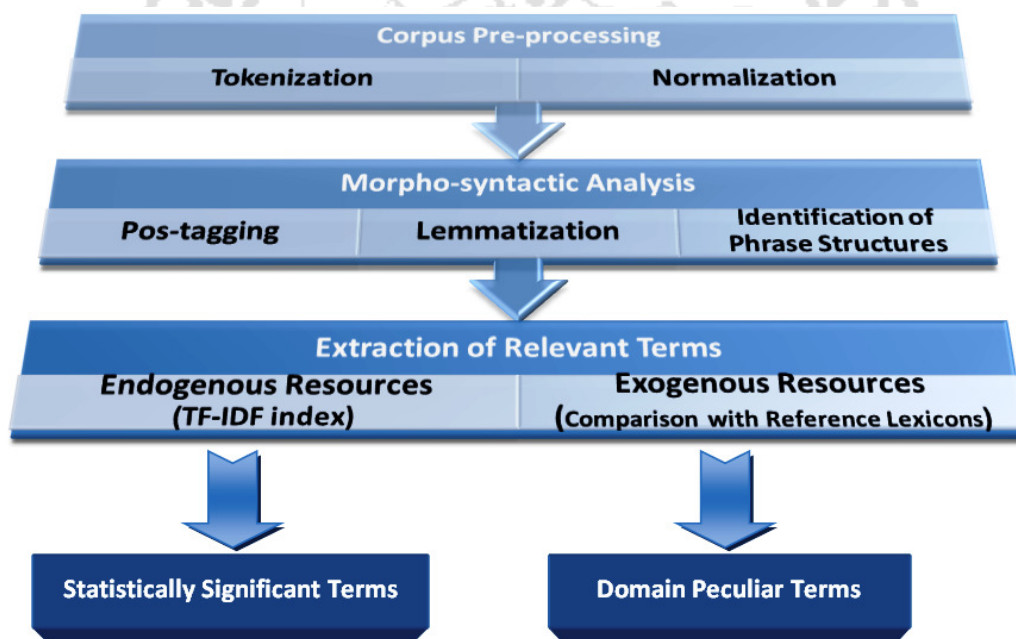


Figure 2.4: general process for extraction of Relevant Terminology

2.5.3 Identification of lexical-semantic relations

Traditionally, knowledge is intended as information about the object surrounding us. These objects can be concrete or abstract entities, properties or relations.

Objects sharing the same properties or characteristics can be grouped into classes from which it is possible to create by abstraction units of knowledge, named concepts. Concepts differ for their characteristics, or semantic traits. Concepts, to have an accurate description, must be related to other concepts, which can be coordinated or subordinate. Since concepts are designated by terms, it is necessary to verify if a relation of sense between terms exists.

The knowledge pertaining to a specialist domain can be organized in a conceptual system by means of hierarchical and non-hierarchical relations between concepts.

Among hierarchical relations are hyponymy and hyperonymy, which describe the relation between a term designating a subordinate concept (hyponym) and a term designating a superordinate concept (hyperonym). This kind of relations is useful to create a conceptual taxonomy. Among non-hierarchical relations is, instead, synonymy: terms designating the same concept are defined synonyms. This kind of relation is useful to create synsets, that is a set of semantic similar terms.

Concerning semantic relation (SR) extraction, it is possible to classify current

approaches in three groups:

Systems based on the distributional properties of words. These systems are based on the distributive hypothesis of Harris' (1968): they consist, in fact, in analyzing the distributions of words in order to compute a semantic distance between the concepts represented by those words. This distance can be used, for example, for hierarchical clustering to automatically derive hierarchies of concepts from texts (Faure et alii 1998; Lee 1997), for Formal Concept Analysis (Cimiano and Staab 2004), for the classification of words inside existing ontologies (Alfonseca and Manandhar 2002; Pekar and Staab 2003) and to learn concept hierarchies (Caraballo 1999; Widdows 2003). Maedche and Staab (2000) and Gasperin et alii (2002) learn association rules from syntactic dependencies between words which, combined with heuristics, are used to extract non-taxonomic relations.

Systems based on pattern extraction and matching. They rely on lexico-syntactic patterns to discover semantic relations between words in unrestricted texts. Hearst (1992) pioneered using patterns to extract hypernymy relations; Berland and Charniak (1999) applied the same technique to extract meronymy. More recently Girju et alii (2006) have studied meronymic relations extraction while Turney (2008) has proposed a uniform approach for the extraction of different kinds of relations from text. Several techniques aim at providing support for the automatic (or semi-automatic) definition of the patterns to be used for SR extraction. Hearst (1998) proposes to look for co-occurrences of word pairs

appearing in a specific relation inside WordNet. Turney (2006) presents an unsupervised learning algorithm that mines large text corpora for patterns expressing implicit semantic relations.

Hybrid approaches. They combine statistical and pattern-based techniques, as in Alfonseca and Manandhar (2002) that have extended WordNet with concepts extracted from *The Lord of the Rings*. Cederberg and Widdows (2003) have applied Latent Semantic Analysis to improve pattern-based hyponymy relations learning. More recently, Ryu and Choi (2007) have proposed an algorithm for IS-A relation extraction from the English Wikipedia. Giovannetti et alii (2008) propose a methodology that integrates lexico-syntactic patterns, manually defined, (pattern-based approach) and a distributionally-based algorithm (statistical approach) to look for instances of the relations of hyponymy, meronymy, co-hyponymy and near-synonymy from a part of the Italian Wikipedia. Lame (2005) performs a syntactical analysis combined with a statistical analysis to look for syntactic dependencies and semantically related words within a corpus of legal documents.

The strategies to extract semantic information from corpora can also be divided into two categories:

Knowledge-rich methods. They require some sort of previously encoded semantic information such as domain-dependent knowledge structures, semantic tagged training corpora, semantic resources like thesauri and dictionaries. How-

ever, this approach inherits the limitations of external resources, like limited vocabulary size, since they can include general words and not the necessary domain-specific ones.

Knowledge-poor methods. They use no presupposed semantic knowledge but try to automatically extract semantic information by observing the various syntactic contexts. In particular, they attempt to extract the frequency of co-occurrence of words within the various contexts to compute semantic similarity among words. The syntactic-based strategy requires specific linguistic information such as assignment of a morpho-syntactic category to each word of the corpus at issue, identification of relevant phrasal structures, identification of syntactic functions, etc. Each word of the corpus is, then, associated to a set of syntactic contexts: words sharing a great number of contexts are considered as similar (Agustini et alii 2001).

Creation of synsets: synonyms

There is unfortunately no neat way to characterize synonyms.

First of all, it is clear that synonyms must have a significant degree of semantic overlap, that is a relevant number of common semantic traits. However, this doesn't mean that the more semantic traits a pair of words share, the more synonymous they are. Consider the following pairs: "animale" vs "albero", "penna" vs "libro", "cane" vs "gatto", "alsaziano" vs "spaniel" ("animal" vs "tree", "pen" vs

“book”, “dog” vs “cat”, “alsatian” vs “spaniel”). As we read the list, the semantic overlap between the pairs increases but it doesn’t become synonymy: “alsaziano” and “spaniel” are not synonyms but only two breeds of dog, so they differ for their inner characteristics. This means that synonyms must not only have a high degree of semantic overlap, they must also have a low degree of contrastiveness. It follows that synonyms are words sharing “central” semantic traits but they differ for their “minor” or “peripheral” traits (Cruse, 1986).

Synonyms can also occur together in certain kinds of sentences, where they are used as explanation, that is to clarify the meaning of another word, like in “E’ stato fatto fuori, ovvero è stato licenziato” (“He was cashiered, that is to say, dismissed”), or in “E’ stato ucciso, o meglio giustiziato”(“He has been killed, or better, executed”).

According to a distributional approach (Harris, 1968), two words are semantically similar on the base of the distributional similarity of the different contexts in which they occur keeping the same truth value.

From all these considerations, it follows by intuition that synonyms have similar meanings but it is also to be noted that within the class of synonyms some words are more synonymous than other. This raise the possibility of a *scale of synonymy* starting from absolute synonymy to zero-synonymy, passing through partial synonymy (Cruse, 1986).

Terms designating the same concept are named ***absolute synonyms***, thus hav-

ing perfect identical meaning, if they are mutually replaceable in *all* their contextual relations without altering their truth value.

It is to note that absolute synonymy is almost rare: it is difficult to find two words having the same identical meaning, since the replacement of one word with the other usually creates different shades of meaning. Consider the pairs “padre/papà” (“father/daddy”), “raffreddore/rinite” (“cold/rhinitis”): the second word of the first pair has a more emphasized emotional value whereas the second word of the second pair is used in a more specialist context.

For these reasons, the notion of *partial synonymy* (or quasi-synonymy) is preferred: in this case, the syntactic distribution of the words at issue coincide only partially. Therefore, two terms are “quasi-synonyms” or “partial synonyms” when they are interchangeable in *some* contexts, in accordance with the linguistic register or the geographical region..

Generally, we can state that synonymy is a *partial overlap of meaning*, it is more a question of semantic similarity rather than identity: two words like “dizionario” (“dictionary”) and “vocabolario” (“vocabulary”) can convey the same meaning in certain contexts, like in the sentence “controllare nel vocabolario/dizionario il significato di obsoleto”¹, but in other sentences they cannot be mutually replaced, like in “il vocabolario di Gadda è ricco di dialettismi” (“Gadda’s vocabu-

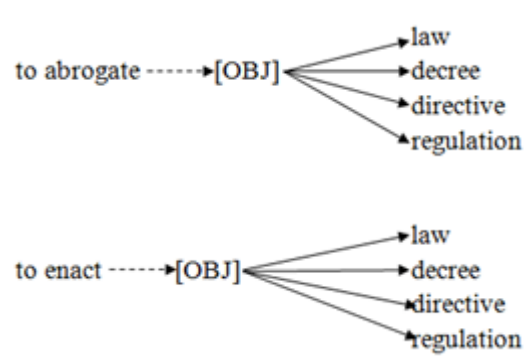
¹In English the words “dictionary” and “vocabulary” are not interchangeable since they convey a completely different meaning. This sentence in English would be “to look up in the dictionary the meaning of the word obsolete”.

lary is full of dialect forms”), where the word “vocabulary” cannot be replaced by “dictionary” (Chiari 2007).

Synonymy enables the creation of synsets, which are classes of words semantically similar. A synset is composed of lexical items belonging to the same part of speech: the terms belonging to the same synset are interchangeable in a context, have the same grammatical behavior and represent different ways to refer to the same concept (that’s why they are also called *variants* of the synset). A variant can be a simple or a complex expression or even an acronym.

The acquisition of groups of semantically similar terms is here performed by taking into account the distribution of the previously extracted terms within the various lexico-syntactic contexts. The distributional properties of the words within a corpus can be, in fact, considered to compute the semantic similarity between the words themselves (Allegrini *et alii*, 2000a, 2000b, 2002 and 2003): according to this approach, two terms are semantically similar if they are distributionally similar, which means that they occur in similar contexts keeping the same syntactic function. This approach identifies a “light” notion of synonymy and denote the presence of a paradigmatic relation between the words at issue: two terms are semantically correlate if they are mutually interchangeable in a significant number of syntactic contexts. To give some concrete examples, the verb “to abrogate” takes the nouns “decree”, “law”, “directive”, “regulation” as complements, and so the verb “to enact”: this suggests that these nouns are semantically

similar since they correlate with the same syntactic function to two verbs.



Obviously, not all contexts are equally relevant to an assessment of semantic similarity between words, that's why the similarity is identified between terms occurring with more selective verbs (with regard to their complements), rather than with less selective verbs: for example, a verb like "to write" is less selective than a verb like "to enact" with respect to the complement "decree" (Dell'Orletta *et al*, 2008).

Creation of a conceptual taxonomy: hyponyms and hyperonyms Hyponymy is the lexical relation corresponding to the inclusion of the meaning of a word into another: a word X is said to be a hyponym of the word Y if X is a (kind of) Y, but not vice versa.

$$X \subset Y \text{ but } Y \not\subset X$$

In other words, the more specific meaning of a word (named *hyponym*) is included into the wider and more general meaning of another word (named *hyperonym* or *superordinate*):

Hyponymy is nothing but a relation of entailment: X is hyponym of Y if X entails Y but is not entailed by Y: *X will be said to be a hyponym of Y (and, by the same token, Y a superordinate of X) if A is f(X) entails but is not entailed by A is f(Y)*” where *f(X)* represents the minimum syntactic elaboration of a lexical item X for it to function as complement of the verb “to be” (Cruse, 1986:88-89)

Therefore, a sentence containing a hyponym unilaterally entails a parallel sentence which is identical in all respects except that it contains a hyperonym in place of the hyponym.

Hyponymy can relate noun to noun, adjective to adjective, as well as verb to verb. In this last case it is also possible to talk about *troponymy*, to show the different nature of the relation between a verb and its superordinate with respect to the one existing between nouns or adjectives.

Here follow some examples: “this is a dog” entails “this is an animal” but “this is an animal” doesn’t entail “this is a dog”; “this is a scarlet skirt” entails but is not entailed by “this is a red skirt”; similarly, “this is the man who was running” entails but is not entailed by “this is the man who was moving”.

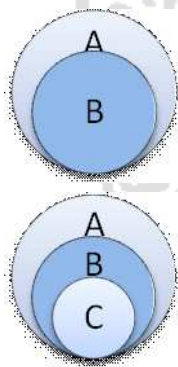
As synonymy can relate the variants belonging to the same synset, hyponymy can relate the variants belonging to different synsets, creating a hierarchical struc-

Dog	HAS_HYPERONYM	animal
Animal	HAS_HYPONYM	dog
To run	HAS_HYPERONYM	to move
To move	HAS_HYPONYM	to run
Scarlet	HAS_HYPERONYM	red
Red	HAS_HYPONYM	scarlet

Table 2.2: Examples of hyponyms and hyperonyms

ture which makes it possible to transfer important semantic information from general to specific concepts, descending to various level of specificity:

$$B \{x_1, x_2, x_3 \dots\} \subset A \{y_1, y_2, y_3 \dots\} \text{ but}$$

$$A \{y_1, y_2, y_3 \dots\} \not\subset B \{x_1, x_2, x_3 \dots\}$$


class B is wholly included in class A

Class C is wholly included in class B and class B is wholly

included in A

Hyponymy is, therefore, a *transitive relation*: if Z is hyponym of X and X is hyponym of Y, then Z is hyponym of Y. In the same way, if class C is subclass of class B and B is subclass of class A, then class C is subclass of class A.

To give a concrete example, “alsaziano”, that is direct hyponym of “cane”, is also hyponym of “animale”.

A principle followed when coding hyponymy is the *principle of economy*: if a word X is hyperonym of a word Y and Y is hyperonym of a word Z, then Z mustn't be directly related to X but to Y. Following the previous example, "alsaziano" must be directly related to "cane" and not to the more general "animale".

This principle serves to avoid that middle nodes in a taxonomy could be left out.

Each hyponym inherits all the properties of the hyperonym, but these properties go to add to the proper characteristics distinguishing the hyponym both from its hyperonym and its co-hyponyms: an alsatian inherits all the properties of the species "dog" (hyperonym) but it has proper characteristics making it specific and different from the other breeds of dog (such as Dalmatian, Pekinese, etc.), which are other hyponyms of dog.

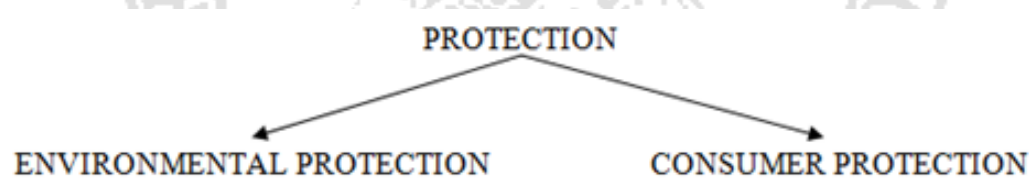
Therefore, X and Y, hyponyms of Z, inherit the general properties of Z, intersect in their semantic traits in common but differ for the traits making them specific.

A pattern-matching approach is used for structuring the previously extracted terms according to hierarchical relations of hyponymy and hyperonymy. These relations are reconstructed from the inner linguistic structure of the text: a complex term is considered as hyponym of another term if this one matches the lexical head of the complex term. To give an example, the complex term "environmental protection" contains the simple term "protection": this helps to deduce that

HYPONYM (more specific term)	HYPERONYM (general term)
ACCORDI CONTRATTUALI	ACCORDI
ACCORDI SINDACALI	ACCORDI
ACQUE CORRENTI	ACQUE
AGEVOLAZIONI FISCALI	AGEVOLAZIONI
...	...
ASSEMBLEA STRAORDINARIA	ASSEMBLEA

Table 2.3: Examples of hyponymic relations extracted from a corpus of legal documents

the concept designated by the complex term “environmental protection” is included in the more general concept designated by the simple term “protection”, consequently, “environmental protection” is hyponym of “protection”. Similarly, “consumer protection” is hyponym of “protection”, and co-hyponym of “environmental protection”, since they share the same lexical head.



Chapter 3

State of the art in Document Management Systems

...TBC...

3.1 Document Management Systems

Starting for the 1980s, a number of vendors began developing systems to manage paper-based documents. These systems managed paper documents, which included not only printed and published documents, but also photos, prints, etc.

Most recently document management systems (DMS) was dedicated at the management of digital documents, this kind of systems commonly provide facilities for document processing as storage, versioning, metadata, security, as well as indexing and retrieval capabilities.

In recent years numerous Document Management projects suitable for specialistic domains is been realized, such kind of system propose funtionality for content Characterization, offering for example, template for the document semi-



Figure 3.1: State of the Art in Commercial Document Management System

automatic generation.

Nowadays DMS are moving toward semantic functionality, including advanced features for contents management as semantic search. A schema of most popular DMS presented on the market, divided for category, is showed in fig 3.1.

In Italy, in the area of specialist domains, numerous projects are presented.

Among the most significant recent experiences, it is worth remembering the FIRB ASTREA Project (Tecnologie dell'informazione e della comunicazione per

la giustizia) realized by the Judicial Systems Research Institute (IRSIG) for the CNR (National Research Centre) in the period 2002-2006. The project, which was developed from the viewpoint of text mining, led to the realization of:

- an automatic document classifier for the categorization of sentences (Giuri-Class), developed by using machine learning techniques and, in particular, the Support Vector Machine;
- a sentence analyzer (GiuriMole), for thematic clustering and the visualization of meta-information, based on the MOLE (Mining On-Line Expert) technology elaborated by CINECA;
- a “legal-metric” analyzer (Giurimetrica), for the extraction of structured information, starting with the gathering of legal documents.

Another text mining strategy, is the TAPA project (Trattamento automatico dei Provvedimenti dell’Antitrust), realized in 2004 for the Anti-trust Authority (Autorit’a Garante della Concorrenza - AGCM). It is comparable, as regards the sphere of applications, to the treatment of legal documents and shows a greater emphasis on statistical and lexico-metrical aspects. The automatic recognition of information on AGCM measures, in fact, was effected using external lexicalisation lists (made available by the Authority itself) followed by the implementation of a series of algorithms based on the recognition of sequences of text with

Regular Expressions (Bolasco et al., 2005). Another relevant experience to be mentioned is the ESTRELLA project (European project for Standardized Transparent Representations in order to Extend Legal Accessibility), financed by the European Union (2006-2008). The main activity of the project was that of developing and validating a standardized open-source platform which enables Public Administrations to define and distribute solutions for knowledge management in the legal sphere. In particular, an exchange protocol was defined for legal knowledge (LKIF), based on standards such as RDF and OWL from the Semantic Web perspective, and the implementation of a platform for interaction with knowledge-based systems in the legal sphere (LKBS), by means of the use of API programming interfaces. As regards the specific notarial domain, it is worth mentioning - taking account of the different set of regulations - the project - X-Not@rial: Sistema de recuperación y Extracción de información notarial, realised by the University of Alicante (Spain 2003), with the aim of knowledge extraction from deeds of purchase , and the NOEMI project (NOtaires Et Minutes), realised by the Centre de Recherche H. Tudor (France 1995 - 2007), with the purpose of publishing on the internet all of the electronic resources relative to documents conserved by the Minutier central des notaires in Paris. The idea of the project is that of a migration of the various formats used over time towards XML using, for the descriptions, international archival rules EAD-EAC and the SDX platform based on Open technologies.

As regards the creation of corpora, the methodological notes contained in the Rapport de Recherche de l'Institut National de Recherche en Informatique et en Automatique (INRIA) "Acquisition et structuration des connaissances en corpus: éléments méthodologiques" (1997) keep their conceptual validity even at the distance of a decade, even though they refer to the specific domain of agriculture. There is little else, apart from a copious number of specific projects - also Italian ones- in which the constitution of corpora has obeyed rules that are not specifically codified or dictated by extemporaneous contingency.

3.2 Multimedia Document Management Systems

Fast access to multimedia information requires the ability to search and organize the information. In such an area the main objective of the researchers is to index in an automatic way multimedia data on the base of their content in order to facilitate and make more effective and efficient the query processing.

In the following, supported by the related state-of-the-art, we describe the major challenges in developing reliable image and text database systems.

3.2.1 Image Database Systems

The goal of an image retrieval system is to find images from an image database while processing a query provided by a user. In the last decade, most of researches are focused on Content Based Image Retrieval (CBIR). The CBIR is characterized

by the ability of a system in retrieving relevant information on the base of image visual content and semantics expressed by means of simple search-attributes or keywords.

Traditionally, CBIR addresses the problem of finding images relevant to the users' information needs from image databases, based principally on low-level image global descriptors (color, texture and shape features) for which automatic extraction methods are available, see [10],[11],[12] for details.

More recently, it has been realized that such global descriptors are not suitable to describe the actual objects within the images and their associated semantics. For these reasons, two main approaches have been proposed to cope with this deficiency: firstly approaches have been developed whereby the image is segmented into multiple regions, and separate descriptors are built for each region; secondly, the use of salient points has been suggested.

Following the first approach, different systems like, SIMPLIcity [13] and Blob-world [14] have been developed. The second approach avoids the problem of segmentation altogether by choosing to describe the image and its contents in a different way. By using salient points or regions within an image, in fact, it is possible to derive a compact image description based around the local attributes of such points [15].

Our proposal [19] follows the second approach avoiding the problem of early segmentation and exploits color, texture and shape features in the principled frame-

work of Animate Vision, according to which is the way that features are dynamically organized in the Where-What space that endows them with information about the context in terms of categories.

The discovered semantic knowledge in terms of categories and relations among them is part of a particular folksonomy produced by humans through the Flickr image management system [21]. It is worth recalling that the use of context/semantics for improving retrieval process is also taken into account by Wang et al. [13], in the form of categories, by Del Bimbo et al. [22], [23], in terms of color-induced sensations in paintings, and clearly addressed by Santini et al. [24], through a mechanism of similarity tuning via relevance feedback. Finally, more recent systems, such as Cortina and ALIPR [25], [26] have as goal the automatic classification of images on the base of low-level features and high-level human annotations.

3.2.2 Text Database Systems

The textual processing phase requires the use of different techniques from interdisciplinary fields: regarding legal ontologies from both theoretical – in order to define legal lexical dictionaries – and application – for organization, storage, retrieval purpose points of view. In order to represent legal knowledge, several works have been proposed, such as: Breuker's Functional Ontology of Law [27], Frame-based Ontology of Visser [28], McCarty's Language of Legal Discourse [29] and Stamper's Norma [30].

As a consequence of such theories, several ontologies are now available, such as Ontology-based Legal Information Environment (ON-LINE), Dutch Unemployment Benefits Act (DUBA) and Cooperative Legal Information Management and Explanation (CLIME).

Several approaches that are based on the wordNet project have been also done: in particular, in Italy, JurWordNet[31] is the first Italian legal ontology. In order to perform identification of concepts and document classification for automatic document description, several works have used pattern recognition techniques, as SCISOR [33] and FASTUS [34].

In the system BREVIDOC, documents are automatically structured and the important sentences are extracted, these sentences are classified according to their relative importance [35]. From the NLP point of view, legal research concentrate on the development of thesauri, machine learning for features recognition, the disambiguation of polysems, automatic clustering and neural networks. The most important systems are FLEXICON, KONTERM, ILAM, RUBRIC, SPIRE, the HYPO extension and SALOMON[29].

3.3 Domain-Document Association

3.3.1 Feature Selection

An Effective method for features extraction from the text is performed by using a graph-based approach that compute Automatic Indexing by Co-occurrence eval-

uation (Ohsawa[43]), that will be used for in our approach. On the basis of the frequency value, a predetermined amount of terms are selected (high frequency set, HF), and added in the initial nodes of the graph. Then is evaluated the association strength between each of these terms using the score function, where of is the occurrence frequency value

$$assoc(term1, term2) = \min(of(term1), of(term2))$$

summed for every sentence in the document. The top $|HF| - 1$ associations are inserted into the graph as edges. If an edge between two terms is the only path that connects them, it is pruned (as depicted in figure 3.2). The graph's

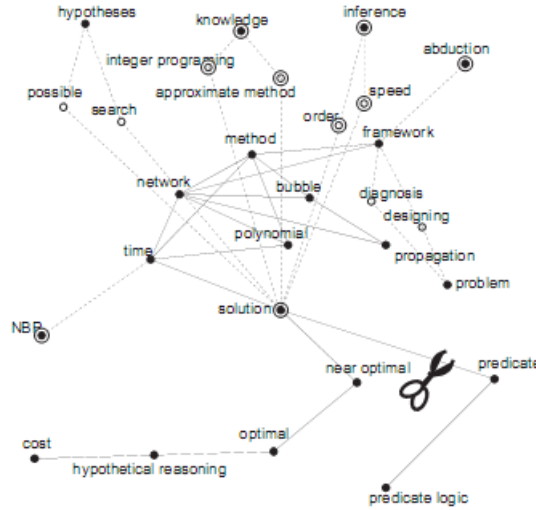


Figure 3.2: Example of scissed weak link

connected subgraphs are then extracted and considered as “concept” clusters. A new batch of terms is added based on their key score, which is the conditional

probability that a term will be used if the author has all the concepts (clusters) in mind ($P(w|g)$) where t is the term and the union is done over every cluster g of the set of clusters. Each of these new terms is then linked to every cluster using the strongest scoring edge amongst the possible ones. Finally, all the terms t in the graph are rated based on the formula that state that $score(t)$ is the summation over every edge connecting t and other terms (w), summation over every sentences s of the document D , of $\min(freq(t), freq(w))$.

$$score(t) = \sum_{\forall e: t \xrightarrow{e} w} \sum_{s \in D} \min(freq(t), freq(w))$$

3.4 Ontology driven human assisted Annotation

The problem of automatically extracting relevant information out of the enormous and steadily growing amount of electronic text data is becoming more and more pressing. To overcome this problem, various technologies for information management systems have been explored within the Natural Language Processing (NLP) and Artificial Intelligence community. Two promising lines of research are represented by the investigation and development of technologies for a) Ontology Learning from document collections, and b) Semantic Annotation of texts. Ontology Learning is concerned with knowledge acquisition from texts as a basis for the construction of ontologies, i.e. an explicit and formal specification of the concepts of a given domain and of the relations holding between them; the

learning process is typically carried out by combining NLP technologies with machine learning techniques. [6] organize the knowledge acquisition process into a "layer cake" of increasingly complex subtasks, ranging from terminology extraction and synonym acquisition to the bootstrapping of concepts and of the relations linking them. Term extraction is a prerequisite for all aspects of ontology learning from text: measures for termhood assessment range from raw frequency to Information Retrieval measures such as TF-IDF, up to more sophisticated measures [10], [8]. The dynamic acquisition of synonyms from texts is typically carried out through clustering techniques as well as lexical associations measures [17], [1]. The most challenging research area in this domain is represented by the identification and extraction of relationships between concepts (taxonomical ones but not only); this research area presents strong connections with the extraction of relational information from texts, both relations and events (see below). Semantic Annotation is the task of automatically identifying in texts instances of semantic classes defined in an ontology [19]. This task includes recognition and semantic classification of items representing the domain referential entities ("Named Entity Recognition" or NER), either "named entities" or any kind of word or expression that refers to a domain specific entity. Recently, annotation of inter-entity relational information is becoming a crucial task: annotated relations range from "place_of", "author_of" etc. to specific events where entities take part in with usually predefined roles ("Relation Extraction"). Currently there exist several SA

systems, addressing different requirements, operating in different domains and on different text types, and extracting different information bits. If we look at the type of SA methodology, systems can be classified into the following classes: - rule-based systems, using hand-crafted annotation rules. Rule-based SA systems are particularly appropriate for dealing with documents showing very regular patterns, such as standard tables of data, Web pages with HTML mark-up, or highly structured text documents such as legislative texts and product catalogues. This is the case of systems like AeroDAML [14], the KIM platform [18], SALEM [3] and PISA [11]; - systems incorporating supervised machine learning: an alternative to the time-consuming process of hand-coding of detailed and specific rules is represented by supervised semantic annotation systems which learn annotation rules from a collection of previously annotated documents. This is the case, for instance, of the MnM annotation tool [20] or of the system developed in the Rainbow project [15]; - systems using unsupervised machine learning: they represent a viable alternative, currently being explored in different SA systems, to supervised machine learning approaches, as they dispense with the need for training data whose production may be as time-consuming as rule hand-coding. Systems based on unsupervised methods can learn from raw text, and for this reason are of great interest. Armadillo [9] and SmartWeb [5] are systems belonging to this category. Depending on nature and depth of the intended interpretation, different amounts of linguistic knowledge must be resorted to. This means that type and role of the

linguistic analysis differ from one SA system to another. The condition part of annotation rules may check the presence of a given lexical item, the syntactic category of words in context and their syntactic dependencies. Different clues such as typographical features, relative position of words, or even coreference relations can also be exploited. Most SA systems therefore involve linguistic text processing and semantic knowledge: segmentation into words, morpho-syntactic tagging, (either shallow or full) syntactic analysis and sometimes even lexical disambiguation, semantic tagging or anaphora resolution. Text analysis can be carried out either at the pre-processing stage or during application of annotation rules. In the former case, the whole text is first analyzed. The analysis is global in the sense that items that are spread all over the document can contribute to build the normalized and enriched representation of the text. Then, the application of annotation rules boils down to a simple filtering process of the enriched representation. In the latter case, text analysis is driven by the process of verifying a rule condition. The analysis is local, focuses on the context of the triggering items of the rules, and fully depends on the conditions to be checked in the selected rules.

3.5 Information Retrieval

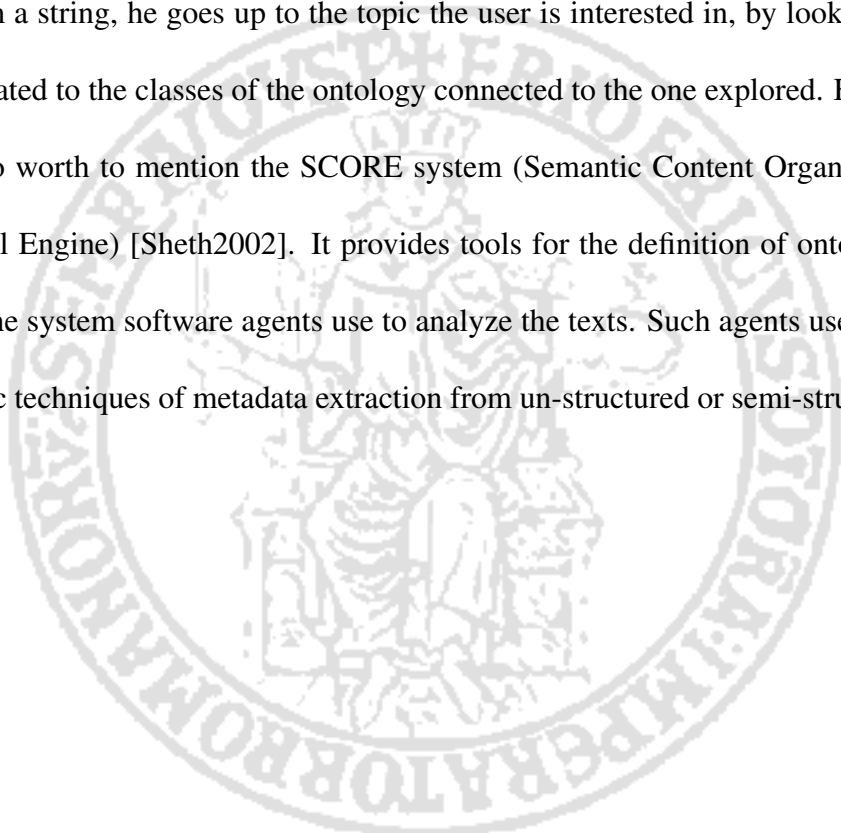
For many years research has been conducted in the field of Information Retrieval (IR) with the aim of allowing machines to automatically retrieve information from different kinds of information sources, among which natural language. Our Unit

will concentrate on Semantic Retrieval, a branch of IR, whose aim is to retrieve semantic content from data sources belonging to the specific domain desired by the user. Semantic Retrieval systems aim at increasing the significance of the retrieved information and they may be classified according to the approach used. The first category is composed of content-based systems. This method, based on the content, is able to collect the preferences of the user and to evaluate the relevance of the pages according to the preferences both of the users and the content. Systems such as Syskill & Webert [Ackerman1997] and WebSail [ChenZ2000] belong to this category. Another approach is domain-knowledge based. It uses both the preferences of the user and the knowledge base; it is organized into domains, in order to improve the relevance of the search results. For instance, Yahoo (<http://www.yahoo.com>) uses this kind of approach, and presents a tassonomic pre-defined path concerning the search made. A typical technique of this approach is the automatic classification of the pages of a taxonomy, both if it is pre-determined or dynamically generated [ChenH2000]. For instance, NorthernLight (<http://www.northernlight.com>) is a kind of search engine which supports the dynamic generation of a taxonomy. By using the “Custom Search Folder” service of NorthernLight, users can refine their query by specifying a domain. This is very useful when the search engine returns excessive information.

Among the approaches used for searching, successful methods have been those based on ontologies. An ontology can be defined as a description of a set of

concepts and the semantic relations existing among the concepts. By using an ontology as a knowledge base, it is possible for an automatic system to "understand" the topic discussed in a web page (topic detection), and to present only the pages related to the semantic domain which has been selected by the user. Currently ontologies for specific domain are being developed for both commercial and public use. Examples of this kind of ontology are: OntoSeek [Guarino1999], On2Broker [Fensel1999], and WebKB [Martin2000]. Ontology-based approaches are very interesting and are widely used in Information Retrieval systems. Here follows the analysis of some IR systems, which currently represent the state of the art in the implementation of such techniques. WebSifter II [Kerschberg2002] integrates a user-centred scheme of evaluation of relevance of the information. The system provides the user with tools in order to generate a taxonomy which is able to represent his specific purpose of the search. Such taxonomy provides the context for the search. The IntelliZap system [Finkelstein2002] is based on the client-server paradigm. A client application sent to the user computer captures the context near the text underlined by the user. The server-based procedures analyze the context, selecting the most important words (eliminating sense- ambiguities) and prepare a set of extended queries for the following search. The basic semantic net is so created through a statistic base and is further enriched by using the linguistic information available on WordNet, an electronic dictionary. Moldovan and Mihalcea system [Moldovan1999] is characterized by the use of an interface in natural lan-

guage, which increases the relevance of the search results. The semantic analysis of the query in natural language allows the expansion of the query submitted to the search engines. An approach based on ontologies and semantic nets is also used by [Picariello2004]. On the base of previous studies on search engines, and starting from a string, he goes up to the topic the user is interested in, by looking for texts related to the classes of the ontology connected to the one explored. Finally, it is also worth to mention the SCORE system (Semantic Content Organization Retrieval Engine) [Sheth2002]. It provides tools for the definition of ontologies which the system software agents use to analyze the texts. Such agents use many semantic techniques of metadata extraction from un-structured or semi-structured texts.



Chapter 4

A Digital Document Model

4.1 A model of document suitable for e-government activity

A document managed in an e-Government information system is usually composed by different multimedia data types, as images, text, graphic objects, audio, video and composite multimedia. This is usually related to two main problems: a multimedia document contains heterogeneous information contents and has to manage different formats. In particular, depending on the authorities which manage the document itself, the same information content is presented in multiple ways, using several presentation formats.

For this reason, in order to opportunely manage and preserve the real useful information contained in a certain document, despite the required different presentation formats, it is necessary to provide a novel model for a multimedia document, pointing out how to:

1. Identify and characterize what is the minimal content of the document itself, given a certain normative context, and
2. Relate this minimal content to a presentation level, depending on different users at different times.

The proposed document model, depicted in figure 4.1 is composed by several layers, as described in the following.

1. *Data Management Layer*: describes the semantic minimal content (or *kernel*) of a document, usually codified by different media types. This layer manages the different data types, furnishing all the necessary functionalities and facilities operating over a certain single media; for example, information extraction and indexing over texts, images, videos, audios and son on.
2. *Integration layer*: provides a proper integration of the heterogeneous data sources, having the aims of regulating the coexistence of the different objects within the context of a single document.
3. *Presentation layer*: this layer regulates the way in which the information has to appear to a single user within a certain context in different times.

In according to such a model, an e-Government document, or more simply

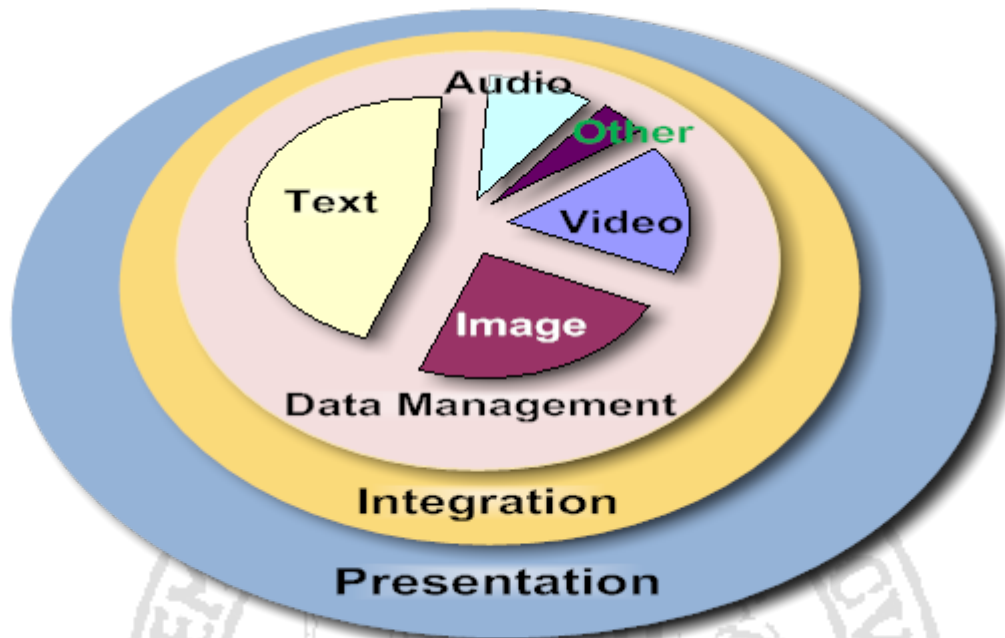


Figure 4.1: The Document Model

e-doc, should be considered as a set of *multimedia assets* that can be opportunely integrated for presentation aims.

From a physical point of view, a multimedia asset is an aggregation of large byte streams, that can be decomposed and represented as a set of structured syntactic components: a text is a sequence of alphanumerical characters that can be organized into words, paragraphs, sections and chapters; an image is a set of pixels that can be grouped into regions; a video is a sequences of frames that can be grouped into shots and scenes; an audio clip is a sequence of audio samples, possibly grouped in audio segments. Then, each kind of multimedia asset has a related precise semantic that describes its content and is necessary for retrieval

and presentation aims.

A generic *multimedia database management system* has to consider both *low-level* (*syntactic*) and *high-level* (*semantic*) features of multimedia objects in order to effectively manage multimedia data.

Thus, a conceptual structure providing semantic information is requested on top of the syntactic representation of raw data, in order to completely characterize multimedia assets and e-docs.

4.1.1 Preliminary Definitions

In this subsection we introduce some preliminary definitions in order to provide a formal definition of the intuitive concept of e-doc from an information retrieval perspective.

Definition 1 [*Multimedia Alphabet*] A MultiMedia-Alphabet (MM-Alphabet) α is a finite set of MM-Symbols ς , where each MM-Symbol is an alphanumeric character or a pixel or an audio sample.

Following the previous definition, two pixels or two characters or two audio samples, i. e. two symbols belonging to the same alphabet, are called *homogeneous* MultiMedia-Symbols. In the case of textual data, a MM-Alphabet is a set of alphanumeric characters. In the case of image data a MM-Alphabet is a set of all possible triples $\langle R, G, B \rangle$, where R , G and B are the color components of a

pixel. Eventually, the MM-Alphabet in the case of audio data is given by a set of audio samples.

Definition 2 [MM-Token] *Given an alphabet α , a MM-Token τ of length k over α is a composition of k homogeneous MM-Symbols from α .*

$$\tau = \langle \varsigma_1, \dots, \varsigma_n \rangle : \varsigma_i \in \alpha, \forall i \in [1, \dots, k]$$

A text or a region of an image are two examples of MM-Token that are composed of a set of alphanumeric characters and pixels respectively.

Definition 3 [MM-Asset] *Given a MM-Alphabet α , a MM-Asset A over α is a composition of MM-Tokens τ , defined over elements of alphabet α , through a set R of relations that represent the logical structure of the asset. $A = (\{\tau\}, R)$. As a particular case, we notice that, if τ is a MM-Token, then $A = (\{\tau\}, \emptyset)$ is still a MM-Asset.*

Definition 4 [MM-Information Source] *A MM-Information Source IS is a set of heterogeneous MM-Assets defined on MM-Alphabets. If k is the cardinality of the asset set, $IS = \wp \left(\bigcup_{i=1}^k A_i \right)$.*

4.1.2 E-Government Document Definition

We are now in the position of introducing the fundamental definition formalizing the concept of *e-Government* document.

Definition 5 [*E-Government Document*] *An E-Gov document is defined as:*

$O = \langle IS, ID, R, l, H \rangle$. Where

1. IS is an element of information source set of MM -Assets composing E-Government document;
2. ID is the set of URIs (Uniform Resource Identifier) of the single MM -Asset;
3. l is a set of low-level relevant features containing a content-based description of all the MM -Tokens (low-level metadata or signature) of component MM -Assets.
4. H is a set of high-level relevant features containing a semantic-based description of all the MM -Tokens (high-level metadata or concepts or semantic description) of component MM -Assets.

An example of the set l for E-Government-Documents containing assets of image and text type is given by visual descriptors coding color, texture and shape of image MM -tokens (whole image and/or decomposed regions) and by classical text features such as number of words, size and format of the document, terms frequency of each asset. The set H may contain semantic descriptors such as a set of relevant keywords, the topic of assets, and so on.

4.2 The RDF Digital Document Model

The core aspect related to a novel and efficient dematerialization process is the idea standing beyond the common concept of document. In Italy, an e-Gov digital document model regulated by recent laws about Public Administration organization.

The starting point of the model is the Document definition of the dpr 445/2000, art. 1, comm. 1, lett. a¹, stating that the representation of the information contained in a document can be unbind from the paper support, and that a document can contain multimedia elements. The proposed model for the bureaucratic document is showed, as *RDF* graph, in fig. 4.2. The three layers, in which the proposed document model is composed, are defined in order to manage and preserve the real useful information contained in the multimedia documents, despite the required different presentation formats. The content will be processed in order to make possible semantic procedure on it, and will be showed in different way, subjected to the Italian normative context, depending on different users at different times .

In appendix A we report the full RDF serialized description of the model depicted in figure4.2, in which the set of documents related of a single thing is en-

¹“Il “documento” è definito come la rappresentazione di atti, fatti e dati su un supporto intelligibile direttamente o attraverso un processo di elaborazione elettronica. Il documento è costituito da oggetti, quali testo, immagini, disegni, dati strutturati, programmi e codici operativi, filmati ed altro che, in base alla loro disposizione sul supporto, ne determinano la forma e, attraverso le relazioni che fra essi sussistono, la struttura.”

veloped in a folder². Every document is memorized in a proper format, chosen on the basis of the authority needs or the available technology (for example, it can be memorized in pdf, doc or odt), and is correlated by property, as the name of the author, the date of creation and change. The access right, indicating who and with which privileges the document may be accessed, are associated to the document itself. The Presentation layer codifies this kind of proprieties, associated to the modality on which the document is presented to the final users. When the documents are submitted to the system preliminary procedure extracts the content of the examined document, such content will be organized in a ordered list of segment. Every segment constitutes a portion of the document and is of a single type of media, then it can be a sequence of words of a text delimited by punctuation mark, an image fragment or an audio stream. The relation between the elements of the same segment are modeled, on the basis of the type of media, in the data management layer. In the case of text segment, the contained words are extracted, and NLP and NER procedure are performed, in order to providing lexical, syntactical and semantical information about them. Based to the particular acception, synonymous sets are individuated for each word, and the proper concept is associated to it, giving in this way the possibility to perform, for example, semantic search operations on the documents. For the other media, as images, audios and videos, low level features are individuated and extracted by apposite procedure

²in Italian we named the collection of documents in digital format as "plico informatico"

realized in the data management layer, and concepts to associate to set of these feature are inferred.

The relations about different segments of the same or different media are codified in the Integration Layer, that contains informations as the reference of a segment of text to an image.

In order to show how the model may be useful for e-Gov applications, let us consider the criminal investigation example described in the introduction.

We note that once we submit the investigation documents to our system, the content is extracted and processed. The proper concepts are the associated to the words presented in the document, so it is possible to perform semantic search on them, for example, searching the profiling details of a person, given a name and surname in input, considering for the research the only person that have a conviction on murder charges on them. Another example is the possibility to highlight the words or the image fragment belonging to a given input concept. Once the relation of different segment are individuated, it is possible correlate them, for example indicating that a text segment is the description of a crime scene represented in a photo, or that a text string constitutes the name of the person that speaks in a particular audio text.

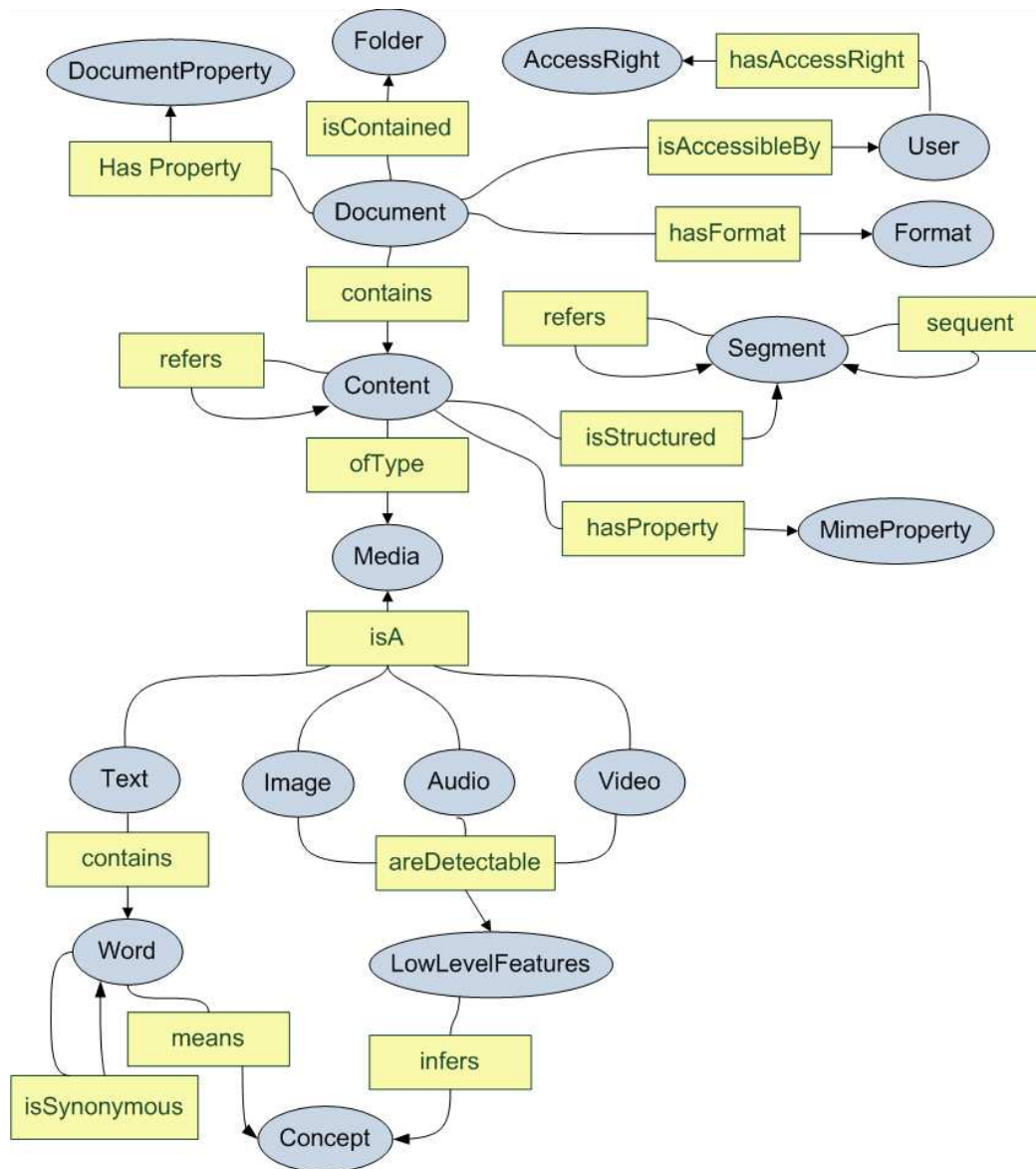


Figure 4.2: Digital Document RDF Model

Chapter 5

An Architecture for Semantic Document Management System

5.1 General Process Overview

In this Paragraph we describe an innovative system of document processing able to accept as input document collections belonging to a specialized domain and to provide automatic procedures for the retrieval of relevant documents, the extraction of relevant information, the presentation of the informative content suitable for the different technologies and the current regulations, and the long term preservation. A schema of such processing is depicted in fig. 5.1.

The belonging of the document collection to the specialized domain represents a desiderata because it allows to considerably reduce the ambiguity resulting from the words interpretation . The whole process of document processing can be divided in three main stages:

1. Domain formalization;

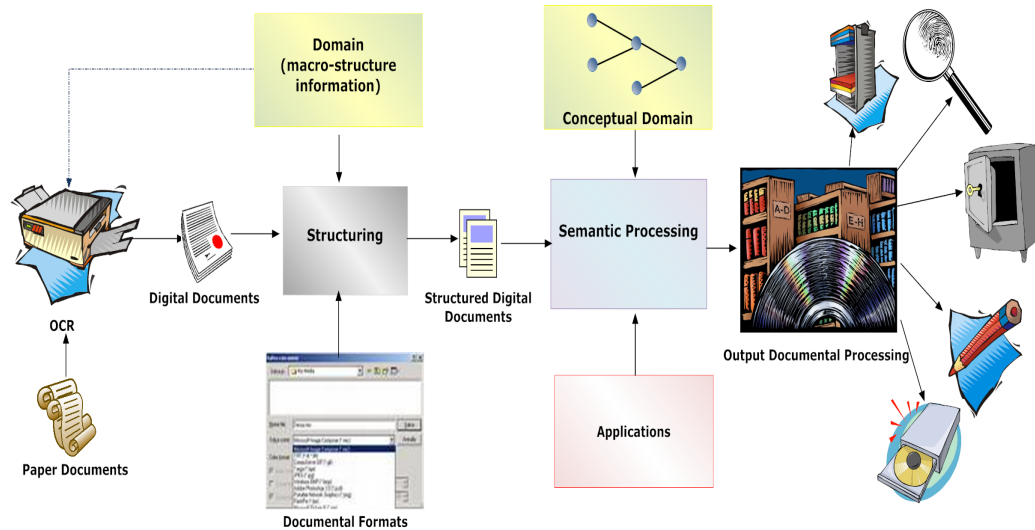


Figure 5.1: General Schema of Whole Documents Processing

2. Document association to the opportune domain of reference;
3. Final users utilization.

These stages characterize the system functioning modes: thanks to its functionalities for document management, the system, in fact, can be also used by unskilled staff to perform automatic operations on documents, such as finding relevant information and performing long term preservation. For these reasons, the third stage is considered as an operating stage. The first stage enables the system configuration and tuning by specialized staff (such as computer engineers and experts in Linguistics) who encode the necessary information for the specification of the relevant data. Ontologies will be used during the whole project as means to encode the information of interest. The second stage concerns the entry of the corpora to process: opportune procedures will guide the user in the choice of the domain

pertaining to the documents submitted. In the following there is a description of the three stages and the respective organization into sub-stages.

Domain formalization This stage aims at encoding with opportune data structures the information of interest pertaining the domain the document belong to. Information is characterized by relevant concepts and relations among them: these elements can be found within the documents to process. This stage is composed of the following sub-stages:

- Extraction of the peculiar lexicon starting from a statistically relevant corpus of documents belonging to the domain to formalize.
- Identification of the relations, of first and second level, occurring among the domain peculiar terms extracted

Document-Domain association Although during the utilization of the system the user can explicitly indicate the domain of reference of the documents submitted, this stage enables the automatic association of the documents to the domain of reference, which is then suggested to the user. This stage involves the use of classification methods aiming at determining the category, that is the domain, the document belong to. This stage provides for the application of:

- Methods of feature extraction

- Document automatic classification by means of well-known methods of Pattern Recognition and Machine Learning

Final users utilization This stage implements the functionalities offered to the user, using the information resulting from the previous stages.

- Indexing procedures for the document search.
- Information Extraction procedures based on:
 - Rule-based Systems
 - Machine Learning
- Procedures to represent information in different formats and according to different access policies.

5.2 The System Architecture

A multimedia database management system is the heart of each multimedia information system such as an e-Government information system: it must support different multimedia data types (e.g. images, text, graphic objects, audio, video, composite multimedia, etc.) plus, in analogy with a traditional DBMS, facilities for the indexing, storage, retrieval, and control of the multimedia data, providing a suitable environment for using and managing multimedia database information.

More in details, a MMDBMS must meet certain special requirements that are usually divided into the following broad categories: multimedia data modeling, huge capacity storage management, information retrieval capabilities, media integration, composition and presentation, multimedia query support, multimedia interface and interactivity, multimedia indexing, high performances and distributed multimedia database management.

All document management system applications should be designed on the top of a MMDBMS in order to support e-Government processes in a more efficient way, in particular those tasks regarding: automatic information extraction from documents, semantic interpretation, storing, long term preservation and retrieval of the extracted information.

The architecture of the proposed MMDBMS system, shown in figure 5.2, can be considered a particular instance of the typical MMDBMS architectural model [38] and is a suitable support for the management of e-Government documents. The main components of the system are the modules delegated to manage the *Information Extraction and Indexing* process and those related to *Retrieval and Presentation* applications. All the knowledge associated to E-Gov documents is managed by apposite *ontology repositories*.

In the current implementation of the system we have realized three main separate subsystems that are responsible of information extraction and presentation tasks: one for the text processing related to e-doc, another one for processing the

other kinds of multimedia information, in particular images, and the last one for presentation aims in according to the requirements of public administrations.

The multimedia indexing and information extraction modules can be also specialized for other kinds of multimedia data such as audio and video. In this case ad-hoc preprocessing components able to effect a *temporal segmentation* of multimedia flow are necessary to efficiently support the indexing process.

The features of text and image management subsystems will be described in the following.

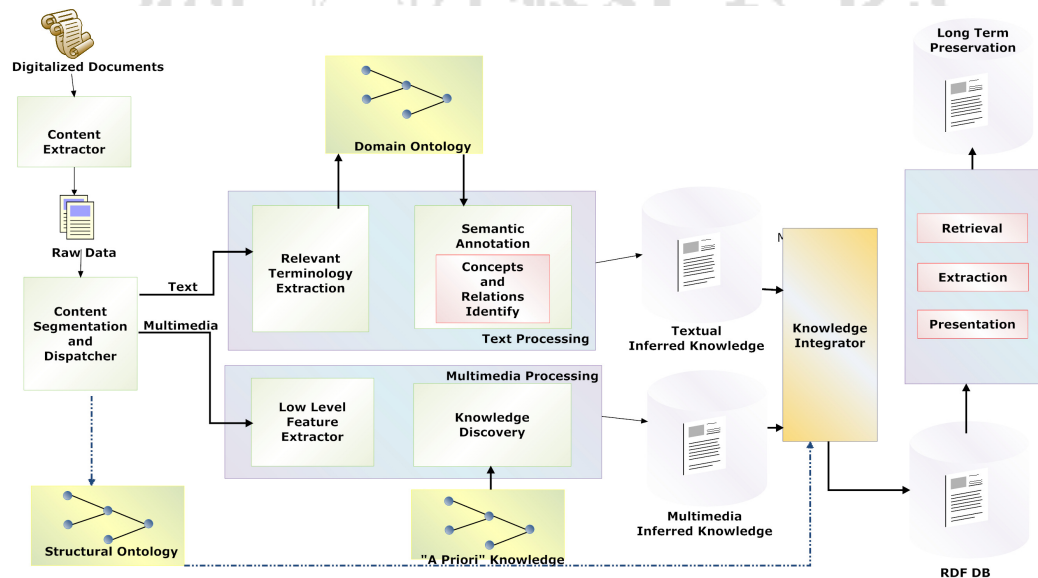


Figure 5.2: System Architecture

5.2.1 The Text Processing Module

The *Text Processing Module* aims at extracting the relevant information from the documents of the E-Government domain, starting from the analysis and the processing of the textual content of the submitted input document.

The defined procedures are based on both linguistic and statistical approaches for the early processing of the submitted input document, together with semantic functions for retrieval and interpretation purposes.

The textual processing methods make use of a knowledge domain, codified by several levels of ontologies, in order to provide the identification and extraction of relevant words in the text, representing the instances of the concept of interest. Such concepts are needed to automatically infer knowledge from data, thus simplifying the information extraction, retrieval and indexing tasks.

For knowledge modeling aims, are defined three main kind of ontologies: (i) lexical ontology, that contains lexicalized concepts commonly used in the Italian and English language, (ii) structural ontology, that codifies the modality in which the information are graphically disposed on the e-Government documents; and (iii) domain ontology, containing the significant and specific concepts and the relations for the interest domain, suitable for the e-Government activity.

The whole processing procedure is composed of several stages [4], [36]:

1. 1))Text extraction, where the plain text is extracted from the source file;

2. 2))Structural analysis, where the textual macrostructures are identified for text sections recognition;
3. 3))Lexical analysis, where each text element is associated with a grammatical category (verb, noun, adjective etc.) and a syntactic role (subject, predicate, complement, etc.);
4. 4))Semantic analysis where proper concepts are associated with discovered entities and relations among them, by means of structural, legal domain, and lexical ontologies. Such procedures make a proper semantic annotation that is codified by RDF triple [39].

5.2.2 The Multimedia Processing Module

The goal of the *Multimedia Processing* subsystem is to automatically infer useful annotations for multimedia data (images) looking at their visual content and exploiting an “a priori knowledge” (obtained in the training step of the system) in the shape of multimedia ontologies[40].

Such ontologies formally represent relationships between raw data features and semantic concepts relevant for the considered domain and are dynamically built by exploiting pre-defined annotations or taxonomies, for example those pro-

vided by Web 2.0 collaborative environments (web folksonomies of Flickr [21])

To such purposes, each image, belonging to a given concept (category) of the a-priori knowledge, undergoes a particular indexing process, where in a first step a low-level description is obtained and then in a second one an apposite indexing structure is created/updated for facilitating the successive retrieval and annotation tasks. To obtain a low-level description of the images, we applied a salient points technique - based on the *Animate Vision paradigm* - that exploits color, texture and shape information associated with those regions of the image that are relevant to human attention (*Focus of Attention*), in order to obtain a compact characterization, namely *Information Path*, that could be used to evaluate the similarity between images, and for indexing issues. An information path can be seen as a particular data structure: $IP = \langle F(ps; \tau s), hb(Fs), \Sigma Fs \rangle$ that contains, for each region $F(ps; \tau s)$ surrounding a given salient point (where ps is the center of the region and τs is the observation time spent by a human to detect the point), the color features in terms of HSV histogram $hb(Fs)$, and the texture features in terms of wavelet covariance signatures ΣFs (see [10] for more details).

Furthermore, on the multidimensional space defined by image information paths and for each predefined category, we define: (i) a particular index, named *BEM Tree (Balanced Expectation Maximization Tree)*, able to efficiently organize images in the feature space and to provide range query capabilities with good

performances and accuracy for large image databases; (ii) a *similarity measure* between different information paths, that is used to rank and refine range query results.

The proposed indexing process can then efficiently support the *Knowledge Discovery* task (i.e. the “category detection” procedure presented in [19]), which aim is to automatically discover by a probabilistic approach concepts of the a-priori domain taxonomy that better reflect the semantics of input images. Thus, the obtained information can be used as useful annotations for each image, in order to infer knowledge about the content of database images, that is represented in the shape of a multimedia ontology (taxonomy concepts + images).

Finally, the inferred knowledge is coded using an extension of RDF language, i.e. the *probabilistic RDF* [37], because the automatically discovered taxonomy concepts for image are subjected to a given uncertainty.

5.2.3 The Integration and Presentation modules

The objectives of the *Integration and Presentation* modules are: from one hand, to merge in a unique “container” the heterogeneous knowledge coming from text and multimedia data, and from the other one, to delivery the content of e-docs in different formats.

In the current implementation of the system the integration module uses a human-assisted semiautomatic approach to instantiate relationships among con-

cepts of the different ontologies. The result of a such process is an ontology that contains all the knowledge related to the e-gov documents.

The presentation module works on the top of such an ontology and exploiting the set of relations about structure of multimedia assets and e-gov documents in order to present and delivery to final users the content of an e-gov document in different ways: printable (e.g ps), portable (e.g pdf) , word processing (e.g. .doc, .odt, .stw, .rtf, .txt, etc...) and web formats (e.g. XML, HTML).

5.2.4 Document Processing

The documental corpus submitted to the system is processed in order to extract the informative content. An appropriate segmentation task is performed in order to extract the different assets : images, video and text. In case of text detected within images, an OCR/ICR system extract the character sequences. Each segment is then stored in the multimedia DB, in this way, each document is then represented by a collection of heterogeneous data.

Our system provides a categorization task that associates a single document to its proper domain. Each category is thus associated to a domain ontology, produced by means of semi-automatic techniques.

The collection of documents belonging to a certain category is analyzed by lexicometric[41] and incremental bootstrapping[42] procedures that extract peculiar concepts and relations among them, in order to be used for the indexing phase,

for semantic retrieval purpose. Such list of concepts and relations is then refined by domain expert in order to be used for the domain ontology production.

To each document category is associated a structural ontology, formalizing the explicit or implicit rules used in bureaucratic domain for the information disposition in document drafting. In other terms, the structural ontology gives information about the section of the document where the concept are expected to be. This information is really precious for the I.E. techniques, since limit the scope of the rule used for the ontology population.

Eventually the documental system is associated to a lexical ontology that contains the general, non-specific concepts of the language¹, that is used to driven the I.E. procedure for the identification of concepts not included in the domain.

For visualization and Long Term Preservation aims, the stored segments are then associated to appropriate presentation mask that regulated the format (in function of the user preferences, the available technology and the company rules) and the associated security policy, producing in this way different view of the document to users whit different preferences or access rights.

5.3 Domain Characterization

This paragraph contains a description of a technique used for semi-automatic extraction of peculiar lexicon (which is a terminological vocabulary representative

¹Italian for our applications.

of the domain of interests), based on the analysis and the processing of a significant collection of documents belonging to the domain under examination. Once the peculiar lexicon has been extracted, it provides the basis for the construction of the domain conceptual system. This system is codified by means of ontology and it represents the starting point for semantic processing of document contents. Relevant concepts identification firstly requires the ability to identify the entities within the text structure which refer to concepts, and in the second place the ability to identify the constraints to which entities are subjected and the properties characterizing them (Dell'Orletta 2008[?]).

A concept can be defined as a mental representation whose definition should ideally include:

1. *an intentional meaning*, defined by the set of intrinsic properties that are necessary and sufficient to characterize concepts and to make it possible to distinguish them from other concepts;
2. *an extensional meaning*, defined by all the referential entities to which intrinsic properties of concepts are applied;
3. *a lexical expression* used to refer to entities to which concepts apply and to refer to concepts themselves.

Among meanings, the more complex to define is the one of intentional meaning of a concept, while the less is the extensional one.

While operating in specialized domains, the extensional meanings of concepts are simple enough to be managed, since lexicons are more specialized and full and more technical in the intentional meanings of domain concepts. During Interpretations of the document contents, which is dependent by authors and readers shared domain competences and knowledge, the process of coding/decoding concepts from the words can be reached without (*or in the worst case, with a reduced*) ambiguity.

The automatic comprehension of text data involves a series of disciplines. Meanings of documents contents come out from complex, and strongly inter-dependent, syntactic, semantic and pragmatic aspects. Therefore, in order to describe a document and to understand its contents, it is necessary to identify not only the single signs in the document, but also the relations among them, firstly at a syntactic and semantic level and, in second place, at pragmatic level. This means that it is necessary to analyze also the relations the signs have with the external context and in general with the domain the documents pertains to.

Here we propose a methodology for semi-automatic derivation of knowledge from texts in natural language pertaining to specialized domains. The methodology and the techniques applied integrate Linguistics with Statistics for

those aspects regarding the *analysis* and the *interpretation* of text data, with the aim of identifying peculiar concepts of the specific domains conveyed within the documents.

Learning knowledge from texts includes a series of tasks, starting from terminology extraction (for the identification of the relevant entities the domain concepts refer to) and leading to more complex ones, like as the identification of taxonomic and non-taxonomic relations, which aims at the individuation of *synsets* and/or conceptual taxonomies.

The activities of document processing and derivation of knowledge from text have as requirement the identification of words. Not all the words, in fact, are useful for characterizing the semantics of a documental corpus: this is the case of grammatical words, for example articles and prepositions, that, even forming the connective tissue of a text, represent “noise” since they are not carriers of meaningful contents.

Thus, let us consider as *peculiar lexicon* the set of relevant lexical items: it contains the most significant and representative key-words which define the contents of the textual fragments and in general, the whole domain whose corpus is a representative sample set.

Once extracted, the peculiar lexicon will provide the basis for the construction of the domain conceptual system enabling semantic processing of the documental contents by working with the meanings of the resources.

Term-extraction involves a series of sub-tasks, described in chapter 2, that affect different levels of analysis:

1. Text pre-processing: tokenization and normalization procedures;
2. Morpho-syntactic analysis: part-of-speech tagging, lemmatization, identification of phrase structures;
3. Relevant terms extraction.

In these steps we pay particular attention to the identification of phrase structures. In our methodology not only simple words but also complex words, which are syntagmatic combinations of terms, contribute to specific domain concepts definitions.

It is common to find sequences of words that are semantically tied and co-occurring regularly, because of their intrinsic sense of words which make them conceptually associated.

These complex lexical expressions, which lead to a complete and autonomous sense, are very frequent when dealing with specialized domains. Phrase structures represent often specializations of more general concepts (like as the Italian expression “imposta di bollo” – duty stamp – that is a specialization of “imposta” – duty -).

Loosing the overall sense of these sequences during text analyses, may lead to lexical item dispersion: for this reason it is necessary to process complex expressions as autonomous units of analysis.

In order to identify the most significant words in a text both linguistic and statistical approaches are used, in a deeply integrated way: the former goes into the linguistic structures of the text by analyzing the meanings of words; the latter, instead, provides quantitative representations of the identified phenomena. In particular, the strategy for the extraction of peculiar lexicons is given by the integration of

1. Endogenous (corpus based) strategies, like as the extraction of the TF-IDF index (Term Frequency Inverse Document Frequency), by which it is possible to extract the most relevant lexical forms, representing the topics of the documents. It is classically used for identifying index terms, and it is based on the principle that, for every document, the most relevant words occur Many times within a single document, but in a small number of the total documents.
2. Exogenous (external) strategies, like as the comparison of the corpus with domains sublanguages (list of words that certainly belong to the issued domain). The comparison is applied for recognition of shared words, and for

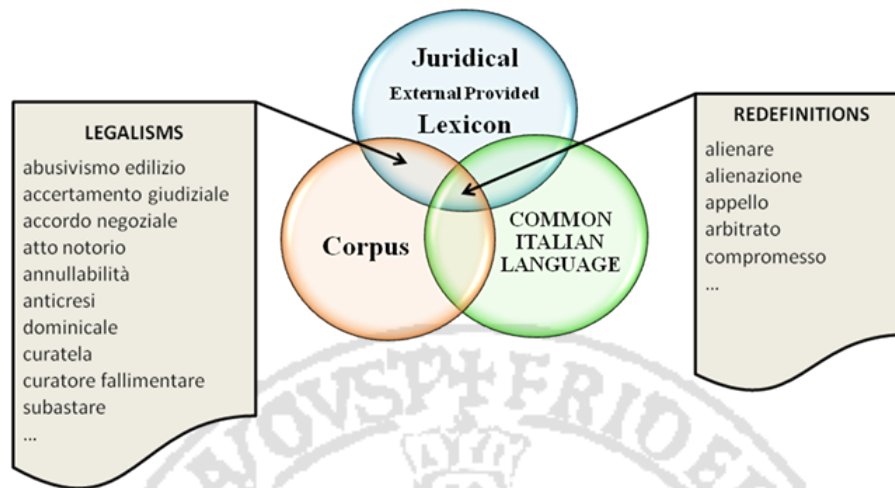


Figure 5.3: Lexical resources to seek legal terms

the identification of the lexical items which are over or under- used with respect to sublanguages of references usually provided by domain experts .

A sublanguage is a specialized language used to provide a definite, technical and precise vocabulary able to cope with the specific needs of a particular domain. For our running example, Law is characterized by its own vocabulary. The vocabulary defines new words or gives other meanings to words already existing in the standard language (this is called *redefinition*). Many terms belonging to general domains, in fact, may be assimilated to legal terms since they label objects, facts or behaviors regulated by law. An example is provided in figure 5.3.

The idea of integration of Statistical and lexical approaches rises from Lame (2005), which has shown that a purely statistical approach produces high values of semantic precision with respect to the corpus contents but poor values of word

recall with respect to the domain language. Statistical indexes, which were classically used to identify index terms, cannot be used to distinguish domain terms from non-domain terms since they do not always correspond with domain terms. Therefore, in order to extract the peculiar words from a document collection with respect to the specific domain of interest, Lame suggests the use of exogenous resources, like as lexical external resources that enable useful comparisons with general or specialized domain terms.

Index terms do not always correspond with **domain terms**. Vice versa, domain terms do not always correspond with lexical items having the highest lexicometric values (for indexing purposes)

In order to define the peculiar lexicon that better represents the domain of interest, our strategy uses a hybrid method, that integrates both linguistic and the statistical approaches. It is based on the Luhn's law (Luhn, 1958) that states that, if we ordered the words in the text by frequency, and considered the distribution of the frequency of the ordered words (fig. 5.4), the index terms between the two cut-offs have the highest discriminating capacity.

We can consider two cut-offs dividing the distribution of the word frequencies into three main sections. The lowest cut-off separates all the words having a high frequency, which are not significant for document characterization (such as generic or common words). On the contrary, the highest cut-off separates rare words which cannot be considered significant enough to be inserted in the peculiar lex-

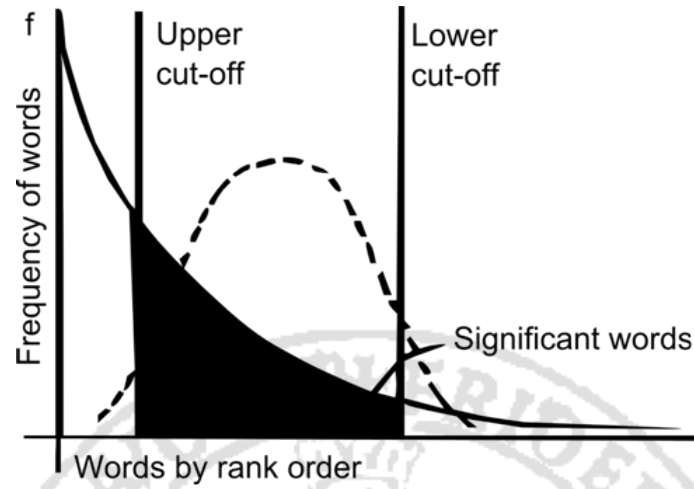


Figure 5.4: Luhn's law

icon, because they are present only in few documents. Conventionally the two cut-offs are set arbitrarily.

Our approach aims at determining the position of the two cut-offs, in order to increase the meaningfulness of the extracted peculiar terms. Such approach, based on endogenous and exogenous extraction strategies, realizes an iterative method that refines cut-off positions depending on the computed distance between the document and lexicon extracted.

The proposed methodology is enacted following the steps depicted in fig. 5.5.

In the first step the *TF-IDF* is computed. In the second step we apply two cut-offs to the index terms list and then the third step the list filtered and the reference vocabulary are compared in order to obtain a temporary peculiar lexical list. In the fourth step the semantic distance among the documents and the temporary peculiar Lexicon is evaluated (using χ^2 measure), and cut-off positions are assessed

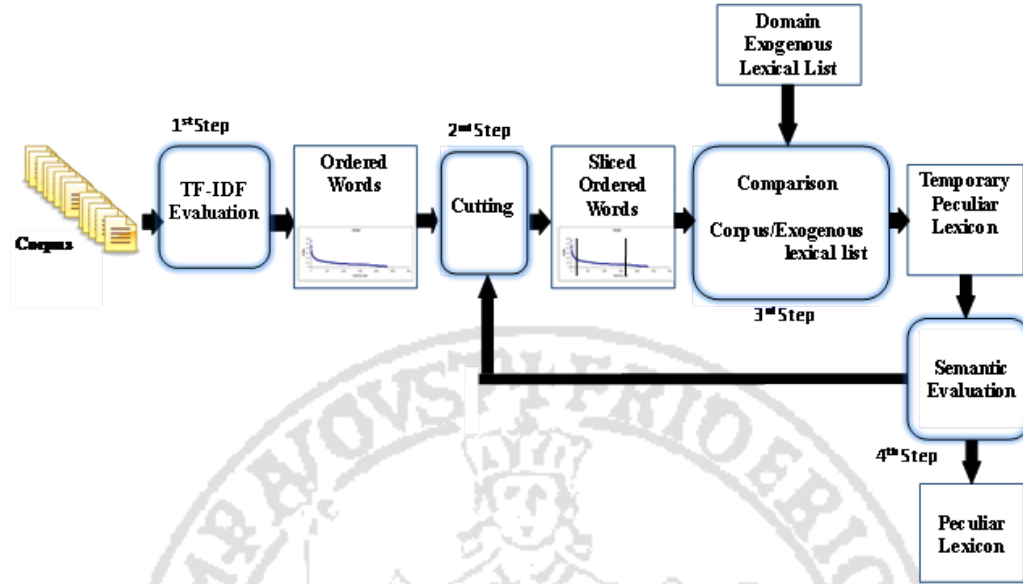


Figure 5.5: Iterative Processing for identification of Peculiar Lexicon

consequently, enlarging the range of selected words if the distance is below some tolerance values, narrowing vice versa.

The evaluation of the semantic distance, in the assessment algorithm devised, is based on:

1. The distance among all the documents, the corpus, the peculiar lexical items (Tab. 5.1);
2. The cover rate of each document and the corpus (Tab. 5.2);
3. The cover rate of each document and the *peculiar lexical items* (Tab. 5.2).

The algorithm is iterated until a satisfying result is obtained (*peculiar lexical items*). For example the similarity analysis performed on a corpus of hetero-

	<i>Doc1</i>	<i>Doc2</i>	<i>Doc3</i>	...	<i>Doc10</i>	...	<i>Corpus</i>	<i>Peculiar lexicon</i>
<i>Doc1</i>	0,00	15,53	16,71	...	17,57	...	15,47	27,25
<i>Doc2</i>	15,53	0,00	3,28	...	4,38	...	2,61	13,18
<i>Doc3</i>	16,71	3,28	0,00	...	5,36	...	3,88	15,15
...								
<i>Corpus</i>	15,47	2,61	3,88	...	4,61	...	0,00	11,70
<i>Peculiarlexicon</i>	27,25	13,18	15,15	...	15,48	...	11,70	0,00

Table 5.1: Chi-squared distance among the documents, the corpus and the peculiar lexical items

	<i>Doc1</i>	<i>Doc2</i>	<i>Doc3</i>	...	<i>Doc10</i>	...
Cover rate respect to corpus	6,022	34,017	19,5	...	16	...
Cover rate respect to lexical peculiar index	2,02	36,364	10,1	...	11,1	...

Table 5.2: Cover rates of each document, the corpus and lexical peculiar index

geneous documents issued by our running example in Notary domain, shows that(5.1), after the first iteration, the document *Doc1* is the worst semantically represented. This is confirmed by the low cover rates in Tab. 5.2. In the same example, the document *Doc2* is instead the best semantically represented.

We execute, therefore, the extraction of a list of relevant words through the TFIDF index and the progressive skimming of the list obtained by comparing it with two different lexicons: firstly a general lexicon for the Italian language and secondly the lexical database of JurWordNet in order to extract a more and more specialized lexicon.

In order to obtain a higher quality of the terms extracted in the document, we have considered the two cut-offs that divide the distribution of the word frequencies into three main sections. The lowest cut-off separates all the words having a high frequency which are not significant for identifying that document (such as generic or common words). On the contrary, the highest cut-off separates rare words which cannot be considered significant to represent that document seman-

tically.

5.4 Domain-Document Association

In order to give a structure to documents of specialized domain, it is possible to divide and organize them into segments by exploiting the information codified in the structural ontology. The same domain ontology, which contains concepts and relations to be extracted from documents, is divided into fragments. Every fragment contains a set of concepts and the relations existing among them. This fragmentation activity is useful for giving a formal objective to information extractions procedures. For example, If for all input documents in a collection segments containing personal data are identified, the information extraction procedures for detecting the name of a person will be performed only on this kind of segments, with a remarkable improvement of precision and efficiency.

As it has been shown in the previous paragraph, in order to associate the proper instances to the ontology fragments, the input documents are segmented in different ways, using several partition rules that are dependent on the specific knowledge domain.

5.4.1 Document Segmentation

In this section we describe the structuring procedure used in our system. Given a document belonging to domain of reference, in order to detect the parts in which it

is structured, turn out a partition of the document. Such partition is performed in order to give structure to the document and further, associate the proper document segment to every ontology fragment.

In order to extract simple fragments of the text we use some partition rules, that are dependent from: *i)* normative prescriptions; *ii)* tradition of single notary schools; *iii)* common use of the single notary. A variety of rules may thus be detected, using several criteria. In the following we give an example of several possible criteria that we have formalized using real notaries expertises.

- Example 5.4.1 (Partition Criteria)**
1. *Starting from the beginning of the document, or from the word that follows the end of the precedent section, every section is ended by the special character ‘.’ followed by ‘\n’.*
 2. *Starting from the beginning of the document, or from the word following the end of the previous section, every section ends before the keywords ‘art.’ or ‘articolo’(law articles in english).*
 3. *To identify each section, we use particular tokens, as “notaio”, “vend”, “acqui”, “compravend”, “rep”, “repertorio”, (in english: notary, sell, buy, article and son on): a section is a portion of text containing one of these tokens. To detect a section, we need to identify its starting and ending word; we thus use the following procedure: let us give three tokens in the document: T_{i-1}, T_i, T_{i+1} , in order to identify the starting word of the section*

relative to T_i , we consider the interval $[T_{i-1}, T_i]$ built using the sequence of words appearing in the document between T_{i-1} and T_i ; we individuate the word w_{middle} located in the middle of this interval. Now we locate the punctuation mark ':' closer to w_{middle} ; if it doesn't appear in the interval, we search for '.', else for ';' or, even, ',', and consider the first word after this. If the interval doesn't contain any punctuation mark, we simply use the w_{middle} word for the section related to T_i . Similar reasoning, on the interval $[T_i, T_{i+1}]$ is done in order to determinate the ending word of the section.

In figure 5.6 we show an example of applying three partition criteria on the same act fragment.

Once several partitions are defined on a given text, we determine the optimal *act partition* on order to associate the most suitable act part to an appropriate ontology module, that contains the concepts and the relations to be extracted.

In order to do that, we apply classification procedure, realized comparing the pattern extracted from each text segment with the concept contained in the ontology module.

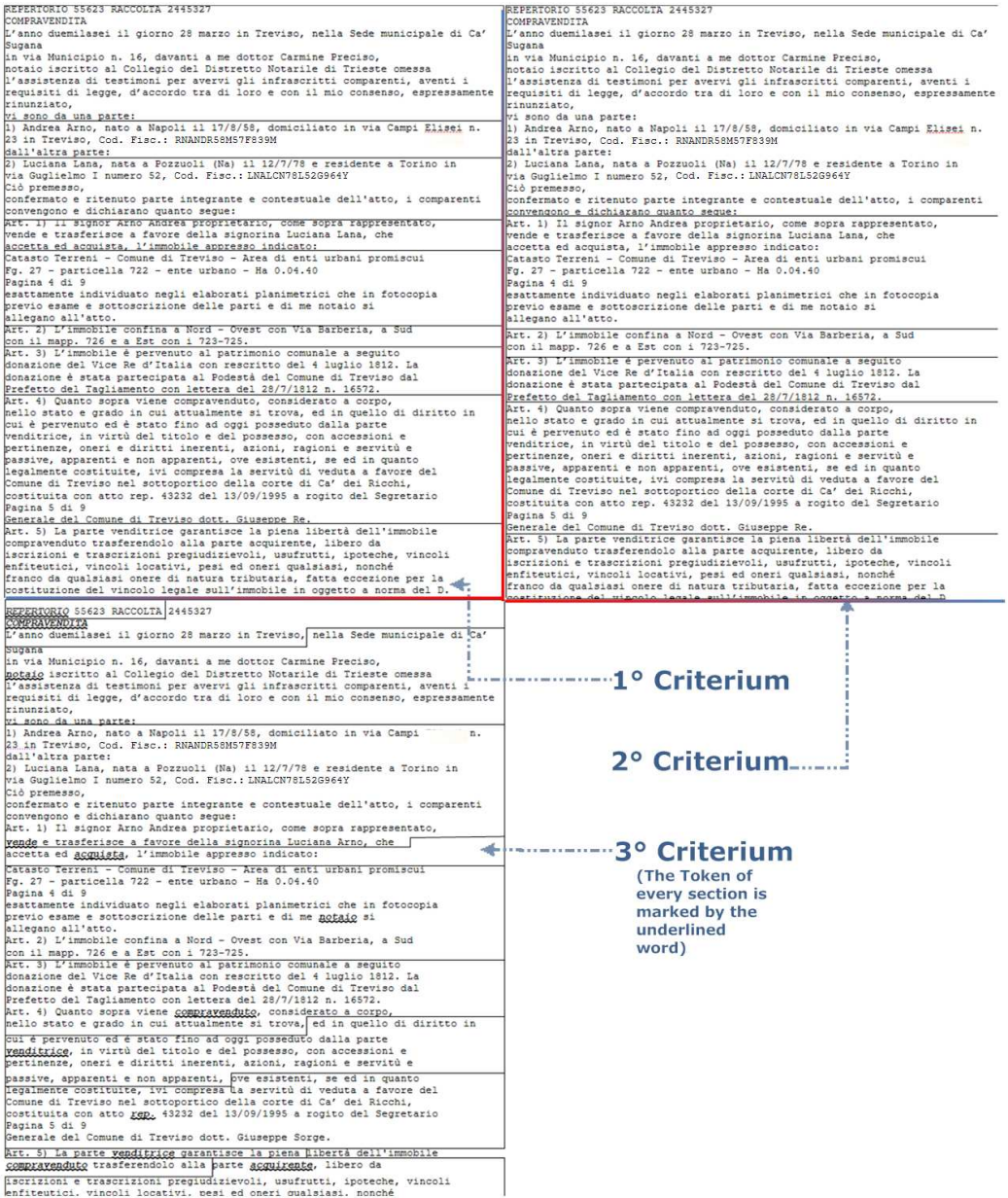


Figure 5.6: Application of tree Partition Criteria on the same Act fragment

5.4.2 Document Classification

In order to give a structure to documents of specialized domain, it is possible to divide and organize them into segments by exploiting the information codified in the structural ontology. The same domain ontology, which contains concepts and relationships to be extracted from documents, is divided into fragments. Every fragment contains a set of concepts and the relationships existing among them. This fragmentation activity is useful for giving a scope to information extractions procedures, thus reducing the text portion in which looking for the desiderata entities. For example, if for all input documents in a collection segments containing personal data are identified, the information extraction procedures for detecting the name of a person will be performed only on this kind of segments, with a remarkable improvement of precision and efficiency.

In order to map text segment to the proper ontology fragment (as showed in fig. 5.7), methodologies of pattern recognition have been exploited.

The structuring document processing is divided in two main workflows: The first is used for the detection of the relevant keywords in the documents, which will constitutes the features for the state of the art classifiers. The second is used for applying the classifiers for identification of the ontology classes to associate to each input document fragment. The latter workflow uses three kinds of classifiers: Naïve Bayes[], Decision Tree[], K-Nearest Neighbor[], the result of classification

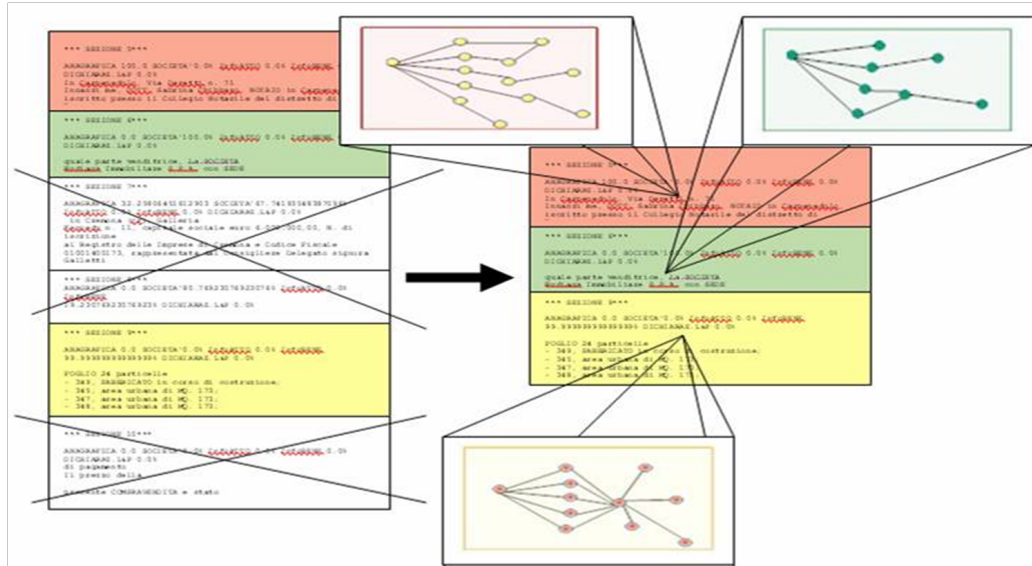


Figure 5.7: Association Document Segments ↔ Ontological Fragments

is combined by using a voting strategy: in case of disagreement, the assigned output class will be the one that get the majority.

Classifiers mentioned above have been chosen because of they implement different classification methodologies and techniques. This is appealing when combining them in a voting methodology since diversity improves results.

The first workflow, aiming to feature extraction, performs textual and natural language preprocessing, enrichment and filtering, together to data manipulation. In particular the textual data of input document segments, is parsed in order to be represented into data structures suitable for the further manipulation. After parsing, the text it is enriched by information about the contained words, obtained by the application of state of the art of linguistic procedure, as the Part of Speech Tagger, which assigns a grammatical category (noun, verb, adjective, adverb, etc.)

to each lexical unit within the input texts collection. The enriched data are then pruned of not meaningful lexical items, cutting out punctuation marks, irrelevant terms (given by stop word list), and non interesting word categories (as article and preposition), detected by POS tagger. This filtered list of terms is then integrated by a list of peculiar domain phrase structures, computed by state of the art of procedures for co-occurrence analysis and analysis of the repeated segments ??.

To quantify the relevance of the elements contained in the resulting list, different term frequencies are computed, and elements are filtered according to these values. To this aim, the term frequency (tf) relative and absolute and the inverse document frequency (idf) are computed, in order to evaluate the tf-idf index for the resulting listed element associated with containing documents. Starting from the selected lexical items, on the basis on their tf-idf index, the more relevant keywords are extracted by means of a graph-based approach that computes Automatic Indexing using Co-occurrence evaluation (Ohsawa[]), described in paragraph 3.3.1. Such keywords will constitutes the features for the classifying tasks. The figure 5.8 shows the workflow for features extraction, that perform the described operations.

The second workflow aims to perform the classification of the input text segments, in order to associate them to the proper ontology fragment, it is showed in fig. 5.9.

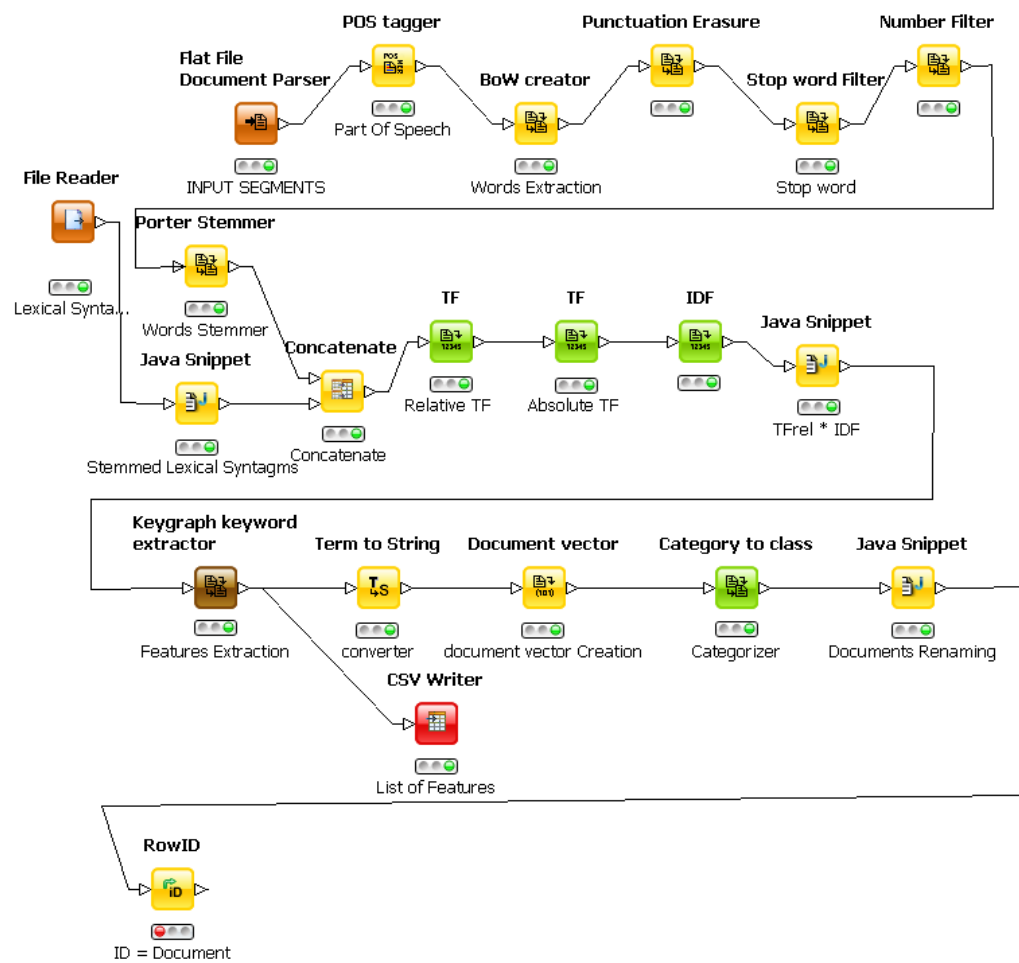


Figure 5.8: Detailed workflow for features extraction

Such list of text segments is classified by using the extracted features. The data to be processed is structured as a matrix, an array of document vector, with a text segment for each row and a feature in each column.

Besides simple manipulations routines on matrices (as the addition of a column for the category tagging) for prearranging the input, allowing it for classifiers processing, the input is submitted to the three selected classifiers: Naïve Bayes[], Decision Tree[], K-Nearest Neighbor[].

The performance of these classification methods are estimated using standard 10-fold cross validation, i.e. the training set is splitted into 10 subsets and every classifier is tested 10 times, using 9 subsets as training set and the remaining subset as testing set. The overall performance is evaluated by averaging the 10 experiments. Such evaluation are used in order to calibrate classifiers parameters. The output of the three classifiers is compared by a voting procedure: each segment is associated to the class indicated by the majority of the classifiers.

5.5 Formal Information Structuring

In the specialized domain almost all the documents is still written using natural languages. Even though, the unstructured form of document follows a well determined sequence: in legal domain, for example in a notary act, the notaries use a certain subset of natural language and in addition they use a certain pre-defined structure, that can be codified by laws or normative rules. For these reasons, we

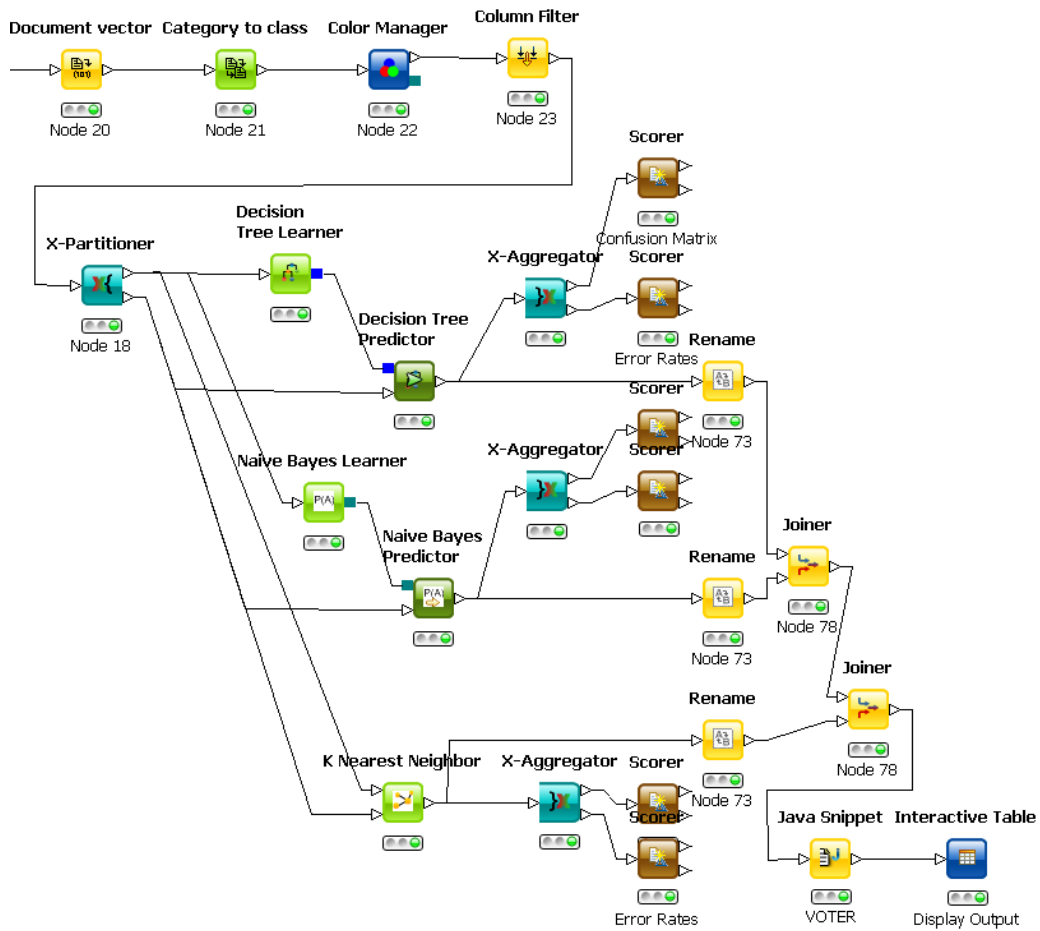


Figure 5.9: Detailed workflow for Segments Classification

say that notaries manage *semi-structured documents* written in a simplified natural language. These considerations are at the basis of the following preliminary definitions aiming to formal structuring the explicit and implicit information that can be detected in a document belonging to a specialized domain.

Structure-UnarySet Let us give a domain \mathcal{D}^S ; a *Structure-UnarySet* (SU) over \mathcal{D}^S is the set of unary predicates, called *structure-concepts* (sc),

$$SU = \{sc_1, \dots, sc_n\}$$

$$sc_i \in \mathcal{D}^S, i \in \{1..n\}$$

Document-Structure-UnarySet A *Document-Structure-UnarySet* (\mathcal{DS}) is a non empty subset of SU containing all the necessary concepts for defining the structure of a given document according to a experts domain description.

Structure-BinarySet Let us give a domain \mathcal{D}^S ; a *Structure-BinarySet* (\mathcal{SR}) over \mathcal{D}^S is the set of binary predicates, called *structure-relations* (sr),

$$\mathcal{SR} = \{sr_1, \dots, sr_m\}$$

$$sr_i \in \mathcal{D}^S, i \in \{1..m\}.$$

Example 5.5.1 (Structures example) According to definition 5.5, a possible SU for the italian notary documents considered is: $\{person, component, date, location, organization, article, section, biographical_section, notary_section, buying_act,$

parties_section}; using example ??, according to definition 5.5, \mathcal{DS} can be $\{\text{article, section, biographical_section, notary_section, buying_act, parties_section}\}$, and according to definition 5.5, $\mathcal{SR} = \{\text{has_number_act, is_part_of, is_kind_of, has_name, has_surname, has_section, has_article, has_sold, is_born_at, has_SSN}\}$

The following definition also stands.

Base-Document Let a *Paragraphs-Sections* (S^P) be the set of textual line inside a document. A *Base-Document* (\mathcal{D}^B) is:

$$\mathcal{D}^B = \{S_1^P, \dots, S_m^P\}$$

$$S_i^P \cap S_j^P \supseteq \emptyset, i, j \in \{1..m\} \wedge i \neq j.$$

In other word, a document is a set of overlapping text-areas; note that we can have different \mathcal{D}^B , depending on the different set of partition criteria used.

In order to capture the knowledge about the structure and the content of the document, let us describe the used ontologies, in terms of their intensional level. First we introduce the *TBox-Module* (\mathcal{TM}) that is used to characterize a fragment of a TBox \mathcal{T} :

TBox-Module Let \mathcal{T} be *TBox*, a *TBox-Module* \mathcal{TM} , is a set of axioms χ that in \mathcal{T} are sound and complete [?].

We can now define *Tbox* as a *Structure-TBox*

Structure-TBox A *Structure-TBox* (*ST*) is a finite set of axioms over concepts and roles belonging to \mathcal{SU} and \mathcal{SR} respectively, expressed according to the syntatic rules and the semantic of $\mathcal{SHOIQ}(D_n)$ description logic.

This kind of intensional knowledge takes into account the document's *implicit* structure used from domain experts to write these legal documents. Considering the notary example, a *Structure-TBox* for a *buyingAct*, may be formed by several axioms selected by a domain experts, e.g the “biographical-section” of a given document, that contains concepts and relations describing “name”, “surname” of “person”, “address” and “security social number”, is represented with the following axioms:

$$buying_act \equiv 4has_section.section,$$

$$biographical_section \sqsubseteq section,$$

$$biographical_section \equiv \geq 2has.person,$$

$$person \equiv \exists hasName \sqcap \exists hasSurname \sqcap \exists hasSSN \sqcap \exists is_born_in.city .$$

These are the set of axioms of the *Structure-TBox*, i.e. the *TBox-Module* related to the *biographical_section* of the *buyingAct*. Each *TBox-Module* has to be characterized by means of a proper key.

In particular, at each key is assigned a feature set associated to regular expressions, keywords occurrences, entity recognition, and a related *score* is computed

considering the positive matching in the feature set; we thus use the best score to detect what is the best module that describes the given fragment. In the following, we will give several definitions used to structure the information related to a document.

KnowledgeKey-Function A *KnowledgeKey-Function* (ψ) is an invertible function:

$$\psi: \mathcal{TM} \longrightarrow k \in \mathcal{K}$$

k being a unique key used to identify \mathcal{TM} and \mathcal{K} the set of these keys.

In our notary example, \mathcal{TM} is identified by a key k^* and the related feature is $feat(k^*) = \{CODICE \setminus s * FISCAL E \setminus s * [A - Z 0 - 9 \setminus s], nat[o, a], an_entity_of_type_person\}$; i.e. a mixture of regular expressions and named entity recognition.

We are now in a position to introduce others concepts related to further levels description of a document D .

Structured-Document A *Structured-Document* \mathcal{SD} is a set of 2-tuples:

$$\mathcal{SD} = \{\langle S_1^P, k_1 \rangle, .. \langle S_h^P, k_h \rangle\}.$$

S_i^P , and $k_i \in \mathcal{K}$ $i \in \{1 \dots h\}$ being *Paragraphs-Sections* and a knowledge key (obtained by applying the ψ function to a \mathcal{TM}) respectively.

Note that different \mathcal{TM} (domain, structure, or lexical) may point to the same *Paragraphs-Sections*; so, some tuples in \mathcal{SD} may have the same *Paragraphs-Sections* and different keys. In our vision, the knowledge related to the notary legal domain should be expressed in a *domain ontology*, including a *structural ontology*, together with a *lexical ontology*.

For example, in an italian notary act we could use a specific legal domain ontology built over the top of JurWordNet [46], several ontologies describing the structure of a particular juridic document produced by domain experts, in addition to a lexical ontology based on ItalWordNet [?].

Given these three different kinds of knowledge, i.e. structural, domain and lexical knowledge, we use the first one for text segmentation aims, the second and third ones to infer more specific concepts related to the semantic content of the documents: in particular, the individuals and the keywords extracted from a section are interpreted as concepts and the relative relations are then inferred using both domain and lexical ontology modules.

Eventually, we represent the extensional knowledge contained in each section in which the document is subdivided:

Knowledge-Chunk A *Knowledge-Chunk* (kc) is an *RDF* triple $kc = \langle r, p, a \rangle$, r being a resource name, p being a property name, a being a value.

We now introduce the last level of description of our legal document:

KnowledgeChunk-Document Let D be a document; a

KnowledgeChunk-Document (\mathcal{KC}^D) is:

$$\mathcal{KC}^D \in \{D, kc_1, \dots, kc_l\}$$

$kc_i, i \in \{1..l\}$ being the Knowledge-Chunk and D the related document.

For example for the “buyingAct”, called $ID-Do-01$, we should have three

Knowledge-Chunk:

Example 5.5.2 (Knowledge-Chunk)

$$\begin{aligned} kc_1 &= \langle myxmlns:ID-Do-01, buyingAct:asset, "Immobile" \rangle, \\ kc_2 &= \langle myxmlns:ID-Pe-01, foaf:name, "Ludovico" \rangle, \\ kc_3 &= \langle myxmlns:ID-Pe-01, buyingAct:seller, myxmlns:ID-Do-01 \rangle, \\ \mathcal{KC}^D &= \{ID-Do-01, kc_1, kc_2, kc_3\} \end{aligned}$$

myxmlns, *foaf* and *buyingAct* being predefined *xml* name space.

The defined strategy for document segmentation and classification can now is considered as the implementation of a function (ρ) that associates an element of *Base-Document* to a *SD*:

$$\rho : \mathcal{D}^B \longrightarrow \mathcal{SD}$$

5.6 Information Extraction and Ontology Population

Once associations between document segments and ontology fragments have been determined, we proceed in populating concepts and relationships in the ontology

Algorithm	<i>RDF-Extractor (RDFex)</i> algorithm
Input: DS	DS is the <i>Structured-Document</i> .
Output: \mathcal{KC}^D ,	\mathcal{KC}^D is the <i>KnowledgeChunk-Document</i>
begin	
$\mathcal{KC}^D = \{D\}$	
foreach $\langle S_i^P, k_j \rangle \in \mathcal{SD}$ do	
$\mathcal{SM} = \psi^{-1}(k_j)$,	
$kc = \text{InferenceProcedure}(\mathcal{SM}, S_i^P)$	
$\mathcal{KC}^D = \mathcal{KC}^D \cup kc$	
end	
end	

Figure 5.10: *RDF-Extractor (RDFex)* algorithm

fragment, by adding proper instances detected in document segments. Relevant information are then extracted, document segments are annotated and results are presented in *RDF* triples containing the properties identified in the segments.

Once the *Structured – Document* is obtained, we extract the *knowledge-chunks* from the text as described in algorithm 5.10.

In this algorithm, the *InferenceProcedure* extracts *knowledge-chunks* from text using inference mechanisms and applying rules for the identification of concepts and relationships instances.

A generic rule is formed by a combination of token and syntactical patterns, which codify the expert domain knowledge. In order to derive instances of relevant concepts or relationships, rules exploit:

- Named Entity Recognition (NER) functionality

```

rule: CompraVendi
(
  ((PERSONA) | (PERSONACC) | (SOCIETA)) :venditore
  ({Token}) *
  ({Token.string=="vende" | {Token.string=="vendo" }}
  ({Token}) *
  ((PERSONA) | (PERSONACC) | (SOCIETA)) :compratore
  ({Token}) *
  ({Token.string=="che"}) ?

  ({Token.string=="accetta" | {Token.string=="accettano" | {Token.string=="compra" | {Token.string=="comprano" | {Token.string=="acquista" | {Token.string=="acquistano" }}
) -->
:venditore.Person = {rule = "CompraVendi",
                     class="http://mia.ontologia#VENDITORE"},
:compratore.Person = {rule = "CompraVendi",
                     class="http://mia.ontologia#COMPRATORE"}

```

Instances previously identified by rules or NER application

String Patterns to match

New Instances inserted in the KB by inference

Figure 5.11: Example of Ruled Based Semantic Annotation

- Morpho-Syntactic information obtained from NLP procedures performed in the Lexical Analysis,

eventually using subsumption on *TBox-Module* for deriving more specific concepts.

An example of rule is shown in example ??, where, using a JAPE[44] style grammar, new instances are detected and annotated in documents, on the basis of strings patterns matching and of existing instances previously identified by same rules applications or by NER application.

The reported rules are able to pick up instances of the buyer and seller from a personal data segment of from a buy-selling act.

The detected instances can be shown by using tools like KIM[45], that high-

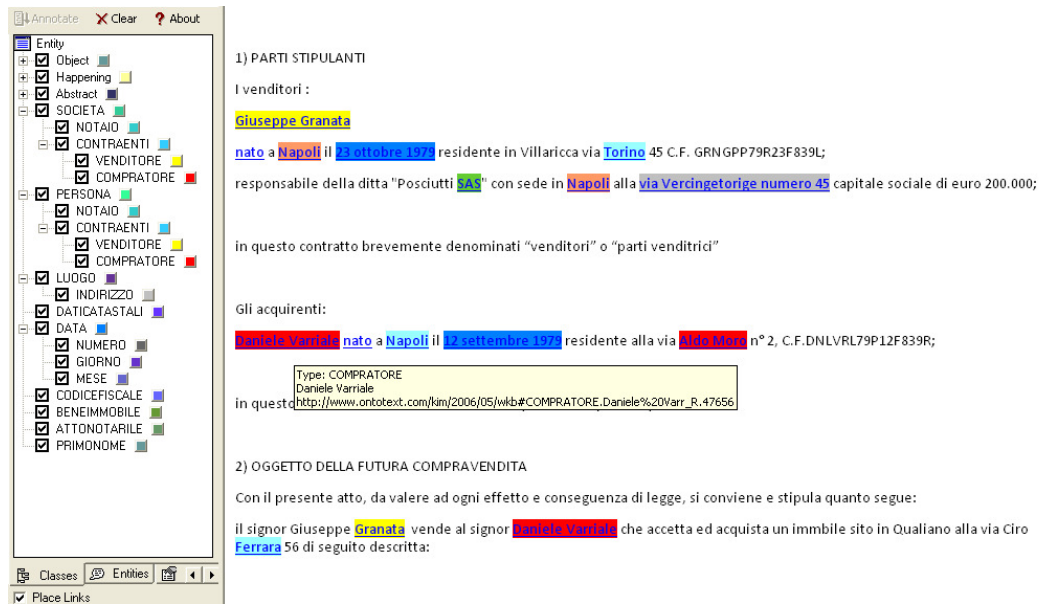


Figure 5.12: An Example of semantic annotation for a Notary Act

lights the associations among detected instances and the concept defined in the domain ontology. An example for the buy-selling act is shown in fig 5.6.

The extracted relevant information is presented in *RDF* triples. For the act reported in the example of figure 5.6, the system extracts several triples from a notary act which defines relationships between the notary and the people involved into the buying-selling process with their generalities. In particular the seller, the buyer, and the relationship among these entities are identified. This is shown in figure 5.6.

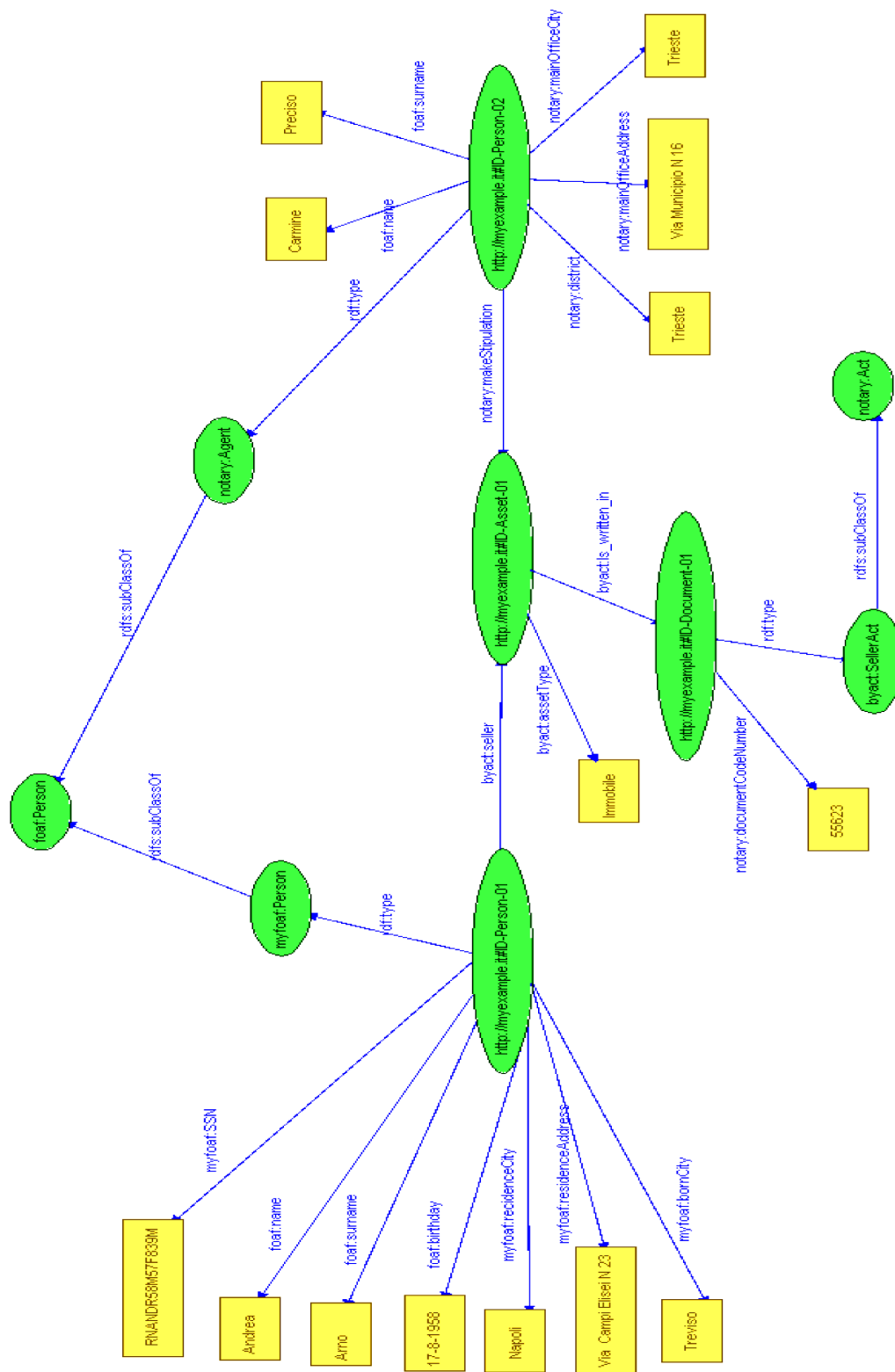


Figure 5.13: A section of *RDF* graph extracted from a Notary Act

5.7 Information Retrieval

Once relevant information related to the domain of interest has been codified for documental corpus, it is possible to execute semantic-based searches which are able to retrieve information by contents and not only by key-words.

The system we propose combines ORDBMS technologies, NLP techniques, proper domain structural ontologies management, and inference rules in order to retrieve significant concepts related to each document and to provide extended querying facilities for users. In particular, one of these facilities is the ability to perform advanced searches that overcome the limit imposed by “keyword-based” traditional queries. It also allows for a “content-based” access to documental database.

Traditional information retrieval systems, based on the comparison of sequences of characters, are in fact able to identify relevant concepts only if they are expressed with the same terms within the text: the search is always limited to the specific key-words inserted into the query and it excludes all the text parts where those keywords do not specifically appear. For instance, when searching for the word “house”, the system will ignore the documents where the words “home” or “residence” appear, even if they represent, in many contexts, the same concept. We exploit, thus, semantic characterization of the document content, in order to improve the quality of the information retrieval. The domain specific knowledge

is represented by means of Ontologies, that contain concepts and relationships among them. Instances of such elements are indicated in the documents by means of semantic annotations, performed by information extractions procedures.

When a search keyword is submitted to the system, the semantic concept it refers to is retrieved. Then all other documents containing terms related to the same concept will be shown as result. The linguistic concepts related to search keywords are represented by means ontologies as *synsets*, which are the set of linguistic nodes related by a synonymy relationship and that can be used in the same statement without modifying its whole meaning. Furthermore, a same term can be used with different meanings. In this case, different synsets are related to different meanings. If these ambiguities are present, the system will provide features to discriminate the synset of interest in the search.

Once these synsets are selected by users, a query expansion mechanism will be used in order to perform queries on corpus where all lemmas in the selected synsets become keywords for a text-based search. The collection of all the documents retrieved from these searches is the results of the semantic-based query. A ranking feature is introduced which scores results depending on tf-idf index which is evaluated for all lemmas too. Notice that all query words and all relevant terms present in documents (which are also used for indexing purposes), have been reduced to their lemma, in order to make the search independent from different declinations and conjugations.

In figure 5.7 a snapshot with the output of the tool used for semantic-based search is reported. In the figure, related to our running example for the legal domain, no particular meaning (in the specialized domain) has been defined for the keyword to search. The search is performed by using all the terms in all synsets related to the keyword. In order to disambiguate the query, in input it is possible to specify the concepts related to the requested search, that will be used to limit the search only to keywords of interest. In the case of the figure, the input lemma for the search is “decreto” (decree), that is present with 8 different meanings in the domain ontology defined for our running example (Instruments for decree , Internal decree, Documentation on decree, decree in Jurisprudence , decree in Legislation, Doctrine on decree, Codes on decree ,Instructions in decree).

5.8 System Description



The proposed *Multimedia Document Management System* to serve its expected purpose has the following main features:


- a unified data model that takes into account content-based and document-based characteristics;
- an ontological support for managing the semantic of data;
- a multi-layer architecture with different kinds of user interfaces;
- advanced functionalities for document indexing and semantic retrieval.



Risultati Ricerca: DECRETO Risultati 1 - 30 su circa 2500 per decreto . Durata della ricerca: 0.07 secondi.


[Istruzioni](#)
[Giurisprudenza](#)
[Documentazione Interna](#)
[Dottrina](#)
[Legislazione](#)
[Ordina per data](#)
[Ordina per importanza](#)



Pagina dei risultati 1 [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [Successive](#)

Istruzioni - **CIRCOLARE** - 22/01/1999 -  - 

1.  **DECRETO LEGISLATIVO 4 DICEMBRE 1997 N. 460, CONCERNENTE IL RIORDINO DELLA DISCIPLINA TRIBUTARIA DEGLI ENTI NON COMMERCIALI E DELLE ONLUS. (sommario non disponibile) DECRETO LEGISLATIVO 4 DICEMBRE 1997 N. 460, CONCERNENTE IL RIORDINO DELLA DISCIPLINA TRIBUTARIA DEGLI ENTI NON COMMERCIALI E DELLE ONLUS. ...**

Giurisprudenza - **TRIBUNALE** - 09/06/1997 -  - 

2.  **IMPRENDITORE INDIVIDUALE - ISCRIZIONE - Presentazione titolo subentro in | precedente attivit?? - Necessit?? - Sussistenza - Comodato di azienda - Forma... (sommario non disponibile) ... Decreto. ... 2556 cc, bensì un vero e proprio comodato di azienda, come risulta ampiamente dalle clausole contrattuali evidenziate nel decreto de quo. ...**

Giurisprudenza - **CASSAZIONE A SEZIONI UNITE** - 25/06/2002 -  - 


3.  **SOCIETA - DI CAPITALI - SOCIET?? PER AZIONI - SCIOGLIMENTO - LIQUIDAZIONE - LIQUIDATORI - NOMINA - Decreto del presidente del tribunale - Atto di vol... (sommario non disponibile) MASSIMA CED: Il decreto con il quale il presidente del tribunale**

Figure 5.14: a snapshot of information retrieval procedure

Figure 5.15 shows at glance the architecture of our system.

Resources in the system are *Digital Documents* (DD) that are managed by a dedicated component, named *Digital Document Repository* (DDR). Its objectives are, from one hand, to allow interoperability among the different data formats by providing import/export procedures and, from the other one, to manage security in the data access. Moreover, documents can be organized in specific *folders* to facilitate the management and retrieval.

In according to the introduced data model, it is possible to associate with a digital document a set of *semantic concepts* – retrievable by semi-automatic information extraction procedures and related to single content units of a document – and set of *keywords* – defined as particular properties of the entire document.

In the early stage, documents acquired by means of apposite OCR techniques are stored in the DDR and undergo the information extraction processing described in the following.

In the indexing stage, digital documents are picked up from DDR by a particular module called *Knowledge Discovery System* (KDS). The KDS analyzes digital documents with the goal of obtaining useful knowledge from raw data. In particular, a *Content Unit Extractor* has the task of extracting (by a human-assisted process) content units from a document (and of generating an instance that can be stored in the system knowledge base), while, the *Multimedia Information Processor* sub-module infers knowledge in terms of semantic concepts from the different

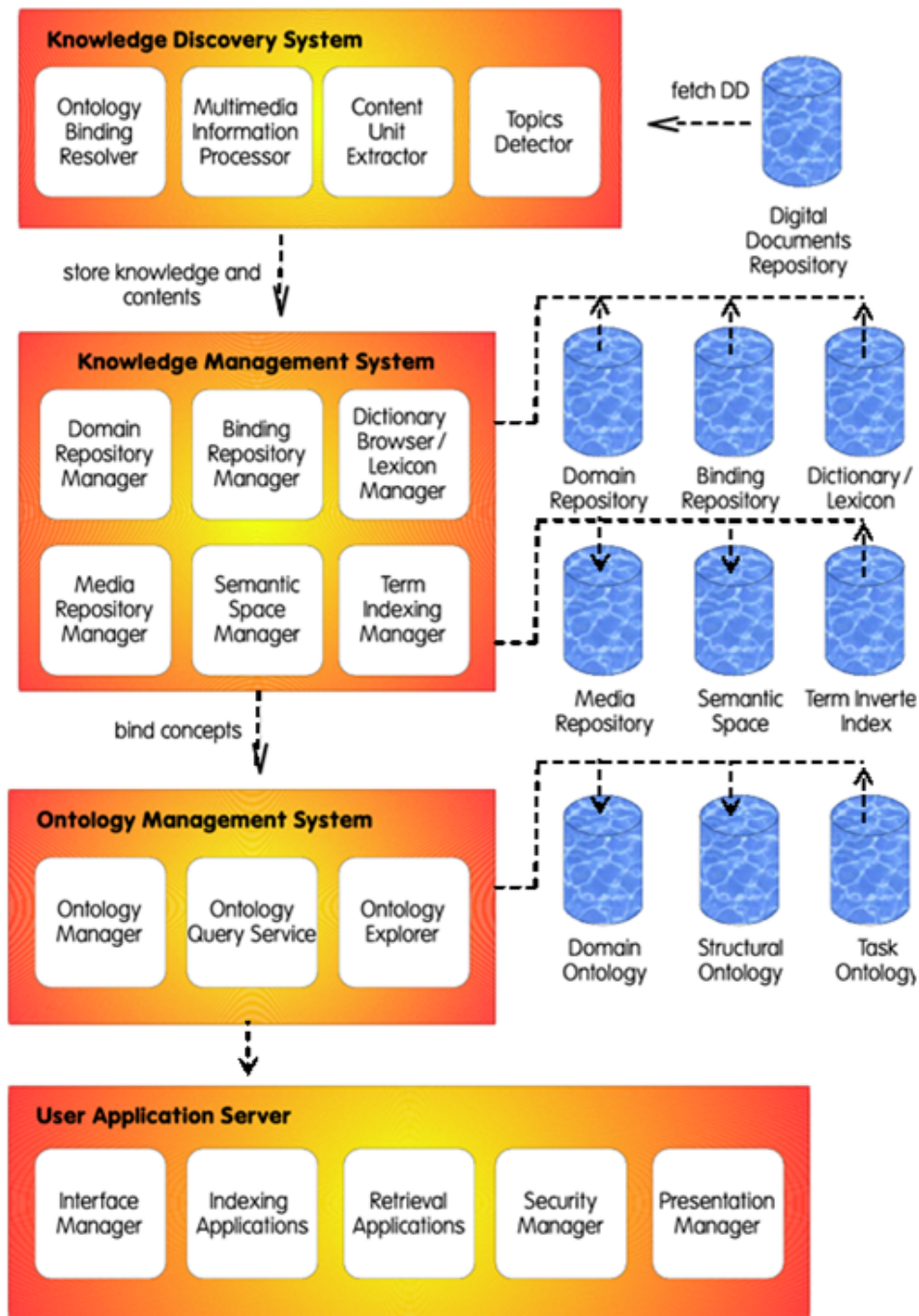


Figure 5.15: System Architecture

kinds of multimedia data [?],[?] (e.g. text, audio, video, image). In the opposite, a *Topics Detector* sub-module operates on the not-structured view of a document and aims at detecting by a natural language processing the most relevant topics for the entire document. Eventually, the *Ontology Binding Resolver* sub-module has the objective of creating for each discovered concept/topic a *binding association* with a node of domain ontology.

The extracted knowledge is then stored in the *Semantic Knowledge Base* (SKB) managed by a *Knowledge Management System* (KMS). The KMS performs indexing operations on the managed information, providing to applications functionalities for browsing and retrieval documents. The components of the SKB (and the related KMS managing modules) are described in the following.

- *Dictionary* (for each supported language) - It contains all the terms of a given language with the related possible meanings and some linguistic relationship among terms (e.g. WordNet). Each dictionary is managed by an apposite management module, called *Dictionary Browser*.
- *Lexicon* - It contains all the terms known by the system: dictionary terms and named entities (names of people and organizations). The is managed by an apposite module, called *Lexicon Manager*.
- *Term Inverted Index* - It is the data structure used for indexing terms inside documents. For each term known by the system (and contained in the

lexicon) a *posting list*, that contains identifiers of documents and contents referring to such a term with the related frequency, is created. The inverted index is managed by an apposite *Term Indexing Manager*.

- *Semantic Space* - It allows the storage of the single atomic pieces of knowledge belonging to document content units, and called *document segments*. It is an abstraction of a shared virtual memory space (with read/write methods) by which applications can exchange multimedia data. This space is called semantic because each element is associated with a particular structural ontology that allows to relate segments of the same content unit and content units of different documents. The *Semantic Space Manager* provides functionalities for reading, writing, removing and searching tuples in the space.
- *Domain Repository* - It contains the description of application domain concepts and is managed by a *Domain repository Manager*.
- *Binding Repository* - it contains the associations between document and domain repository concepts and is managed by a *Binding Repository Manager*.
- *Media Repository* - it is an Object Relational DBMS able to manage the different kinds of multimedia contents. It is managed by a particular module,

called *Multimedia Information Manager* able to support classical multimedia query for the different kinds of multimedia data – e.g. *query by example/feature* for images, *query by content/keywords* for images and text, and so on.

The semantic associated to the data contained in the knowledge base is then managed by the *Ontology Management System* (OMS), that contains the ontology models used by the system. In particular, we exploit three kinds of ontologies (managed by an *Ontology Manager*): (i) a set of *domain ontologies* that relate the semantic concepts in a given domain, (ii) a set of *task ontologies* that determine the role/meaning of a content unit in a document and (iii) a set of *structural ontologies* that code the relationships between contents and segments. The *Ontology Explorer* allows browsing of the concepts in the ontologies, while the *Ontology Query Service* is a component devoted to execute queries on the ontologies.

From the user point of view, the functionalities provided by the system are the *indexing* of a document and the *semantic retrieval* of information. The application interfaces are realized both as web services and desktop programs (and managed by an apposite *Interface Manager*). Finally, there are two modules for *security* and *presentation management*.

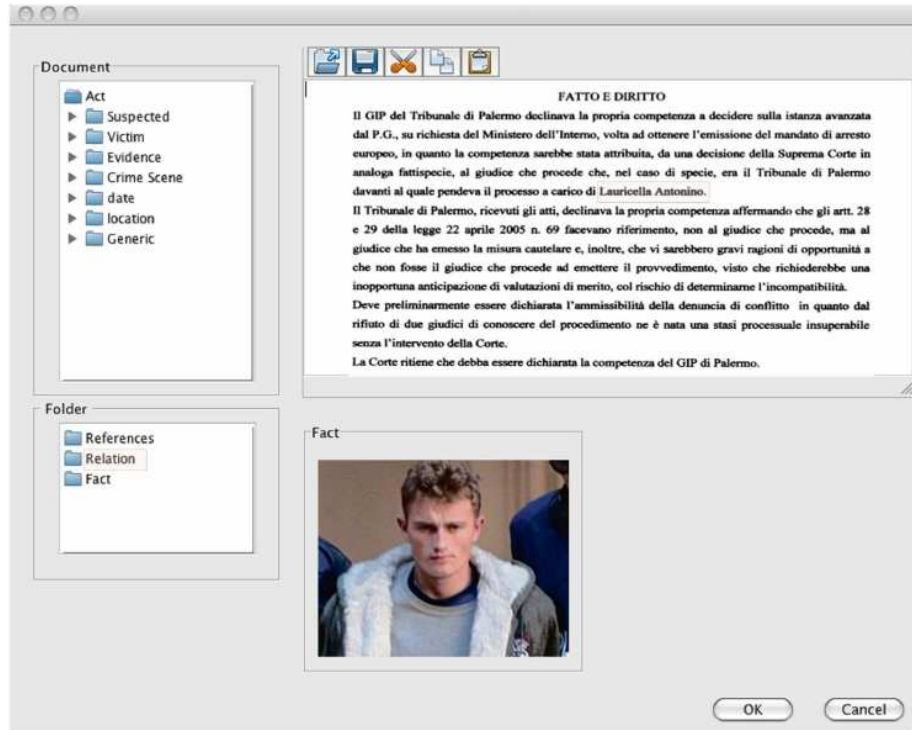


Figure 5.16: Interface for Information Extraction

5.8.1 Implementation Issues

Due to the great amount of data to deal with and security issues, we have chosen to implement the document management system prototype using ORACLE technologies (Oracle 11g DBMS, Oracle Intermedia, Oracle Text, PL/SQL Stored Procedures) for data management and repositories implementation and JAVA both for business and presentation logics.

Oracle Intermedia tools have been exploited, from one hand, to manage images, audio and video stored into the database with the related metadata, and from the other one, to implement the image similarity query. In particular, the oracle

evaluateScore method has been used to implement an image distance through an apposite PL/SQL procedure. Oracle Text functionalities and ad-hoc PL/SQL procedures have been used to manage textual information and implement full-text search.

The ontologies are mapped in the oracle database and managed by the framework KAON 2, while the services of Ontology Query Service are implemented using SPARQL. Eventually, particular JAVA libraries have been exploited to implement Multimedia Information Processing module, Topics Detector, all user interfaces and the other modules.

A couple of interfaces of the prototypical system are presented: in the fig 3 is reported the interface for information extraction features, in which the user is allowed to highlight the relation between a law text under analysis and an image that represent the person to which the content of the text segment references.

In fig. 5.16, is showed the interface that allow the users to submit query to the system².

²To avoid issues with the data privacy legislation, in this work the suspect picture is blurred, in the real system, being the data available to the authorized persons only, the real used images result uncensored.



The screenshot shows a window titled "Find Document:" with a tabbed interface. The "Suspected" tab is selected, with other tabs being "Victim", "Crime Scene", "Evidence", and "Others". The form contains the following fields and controls:

- Identifier:
- First Name:
- Last Name:
- Fiscal Identifier:
- Marital status: (with a dropdown arrow)
- Date of Birth:
- Sex: ☐ male ☐ female
- Pictures: (next to a small photo of a man)
- Match Case: ☐ Whole Words: ☐
- Buttons:

Fig.4: Interface for Information Retrieval

The query are classified on the basis of the subject of interest, that for our domain are: the suspected, the victim, the crime scene and the evidences. In the example the user want retrieve all the acts in which the suspected is the person reported in the pictures that he inserted by the interface.

Chapter 6

Experimental Results for processing of documents in juridical domain

6.1 The Selected Corpus

We have tested the document processing procedure with a documental corpus belonging to legal specialistic domain.

The documents in the corpus have been selected from the Italian Notary Data Base (Banca Dati Notarile). The whole used collection of documents count 66176 documents wich have been produced by the Italian Notary Council (“Consiglio Nazionale del Notariato”).

The corpus is divided into seven sections, each one pertaining to a particular aspect of the normative in acts stipulation. In Fig 6.1 the dimensions, in terms of number of documents, are reported for each section.

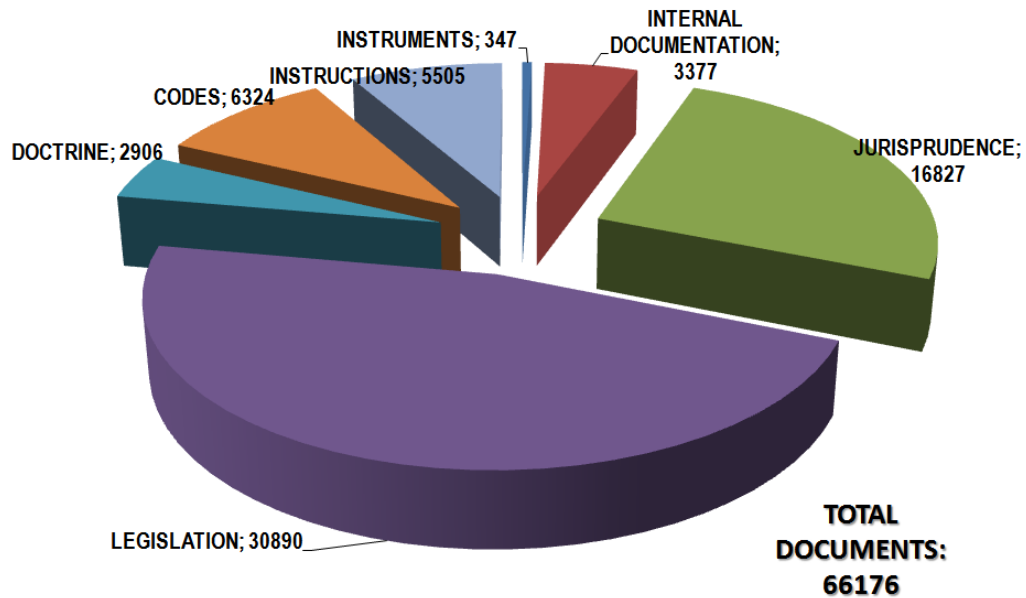


Figure 6.1: the selected corpus

6.2 Domain characterization: Relevant Terms Extraction

The computation of the TF-IDF index has enabled the extraction of the following graphic forms from the analyzed corpus: 203 nouns (out of 837), 36 noun phrases (out of 79), 90 verbs (out of 606) and 1 verb phrase (out of 3) producing a list of 276 lemmas. The list obtained has been firstly compared to JureWordNet[46] lexical database (7768 lemmas) in order to produce an inventory of 160 terms in common, that are the corpus key-words pertaining to the legal domain. Then, this list of 160 lexical items has been further specialized by integration of terms

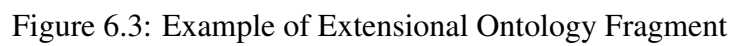
Graphical forms	CAT.	Lemma	TF-IDF	Presence in Reference Lexicon
Società	N	società	7,37981	Yes
Imposta	N	imposta	7,00630	Yes
Notaio	N	notaio	5,65694	Yes
registro	N	registro	5,62684	Yes
questa	Agg	questo	4,74819	No
dichiarazione	N	dichiarazione	4,67598	Yes
donazione	N	donazione	4,32514	Yes
euro	N	euro	4,18925	No
anticresi	N	anticresi	0,02124	Yes
alienabilità	N	alienabilità	0,11230	Yes
alienazione	N	alienazione	2,08754	Yes
arbitrato	N	arbitrato	0,05153	Yes
curatela	N	curatela	0,04720	Yes
abitazione non di lusso	SN	abitazione non di lusso	1,20729	Yes
accettazione esecutiva	SN	accettazione esecutiva	0,00838	Yes
accollo di debiti	SN	accollo di debito	0,07195	Yes
accollo di mutuo	SN	accollo di mutuo	0,03554	Yes
acconti di capitale	SN	acconto di capitale	0,02039	Yes

Figure 6.2: List Fragment of Extracted Relevant Terms

belonging to the notarial domain and identified from the remaining 116 words with no correspondence in JureWordNet. This identification has been performed by eliminating general words in common with a standard lexicon of the Italian language. This has produced a list of 19 corpus specific lemmas that have been integrated to the initial list of 160 lemmas. Here we give a fragment of the list of extracted lemmas in fig 6.2.

6.3 Documents Classification

In this section we report results from executions of classification procedures performed on segments belonging to legal domain. As shown in fig. 1.2 , each segment can be referred to a conceptual area, in which a subset of information can



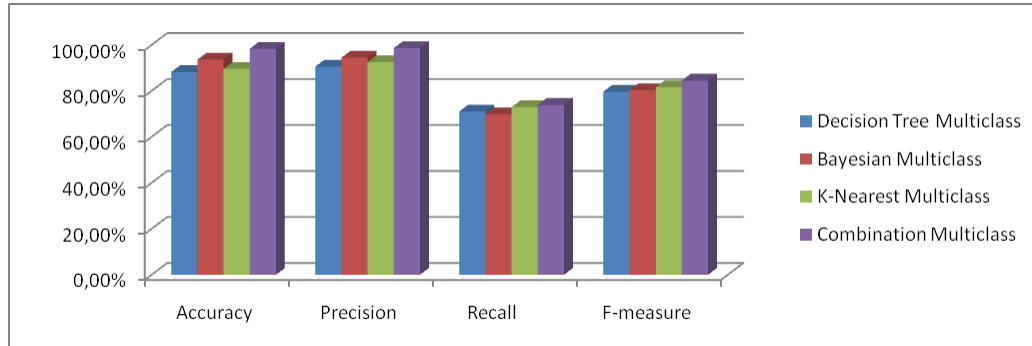


Figure 6.4: Classification Results

be contained. We have classified these segments with the aim of associating the proper class (i.e. the fragment of ontology that contains the concepts instantiates) to the input segment.

We have used three kinds of classifiers, combining (by voting) their outputs and obtaining the results shown in table ???. It is possible to notice the performances improvements of the result obtained from the combined strategy, in respect to the output of single classifiers. Such improvements are due to the diversity of the used classifiers.

The classification results for the voting strategy compared to the best “single” classifier have an improvement of $\sim 5\%$ in precision and $\sim 1,5\%$ in recall, for an overall improvement in accuracy of $\sim 5\%$.

6.4 Information Extraction

In this section we show some experiments we have carried out for evaluating the impact of the proposed system on enhancing user effort in indexing juridical

Precision Value	Number of documents
100%	268
66,6%	23
33,3%	9

Table 6.1: Indexing Precision

documents.

To this aim, the data set is constituted by 350 documents of two different notary schools; 50 documents have been used as training set to train the classifier used for text segmentation.

The objective in this experimentation is to evaluate the system correctness (precision) in automatically discovering the relevant concepts of a juridical document and in particular:

1. the seller with the related personal data;
2. the buyer with the related personal data;
3. the purchase object and its characteristics.

Table 1 shows the related results and in particular the number of documents that has a given value of precision (100%: all the concepts have been correctly discovered, 66.66%: two concepts have been correctly discovered, 33.33%: only one concept has been correctly discovered).

In the majority of cases for which precision is 33.33% or 66.66% the found

Document size	Indexing Times (s)
<50K	1,5
50k – 100k	1,8
100k – 200k	2
>200k	2,5

Table 6.2: Indexing Times

correct relevant concepts are the buyer or seller, thus in our approach the most difficult concept to discovery is that related to purchase object and its characteristics.

Table 2 shows the average indexing times with respect to the document size. Eventually, a snapshot of a Fragment of Extensional Ontology with the extracted instances is reported in fig.fig:ontology.

6.5 Information Retrieval

For Information retrieval evaluation, let us call *top keyword x* TK_x for short, the set of the x more representative keywords for our category (depending on the score associated to each word). For example, TK_{50} will be the set of the first 50 entries in the score-ordered list of representative words for a given category (which also coincide with the top 50 words in Fig.6.5

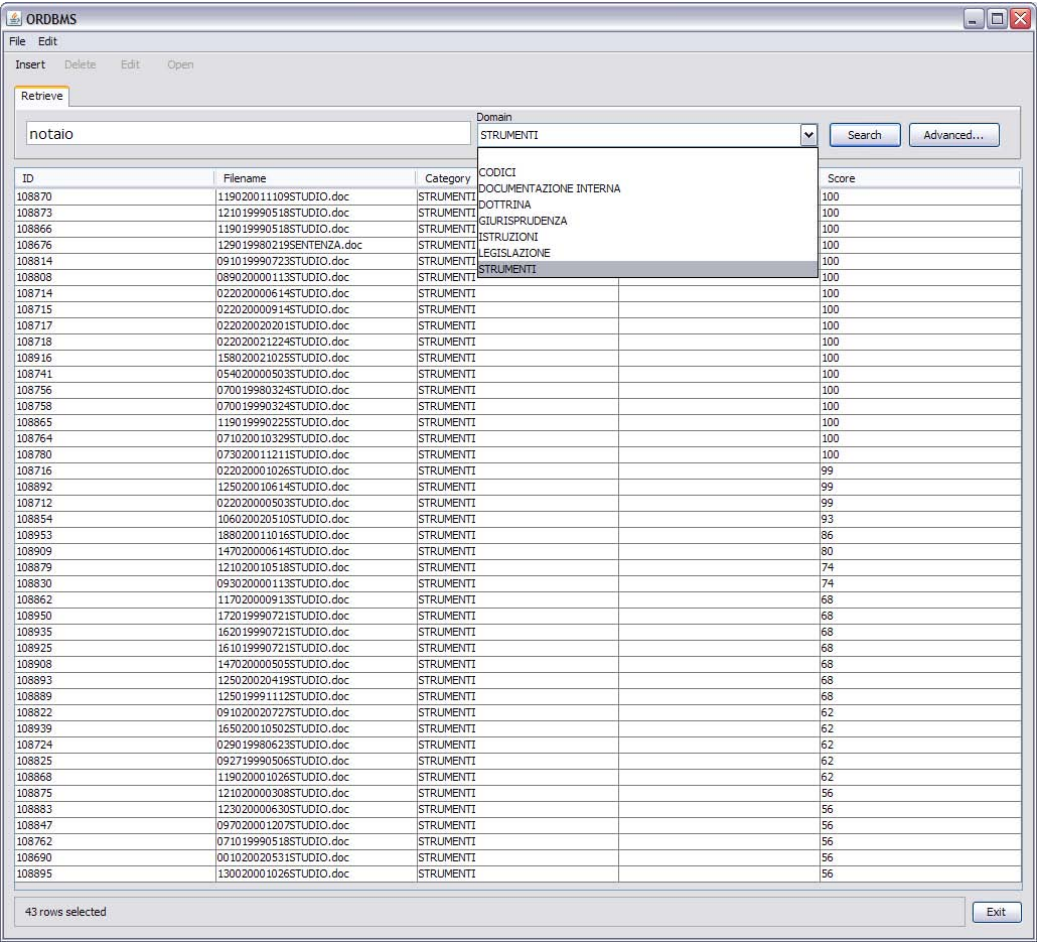


Figure 6.5: *TK* sets selection

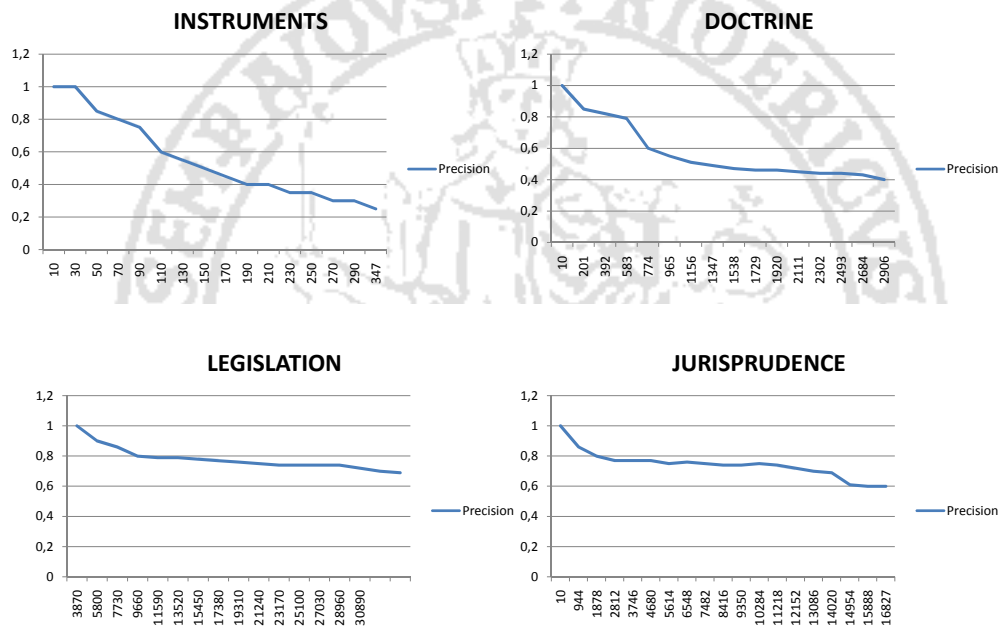


Figure 6.6: Experimental Result for Information Retrieval task for selected categories

Chapter 7

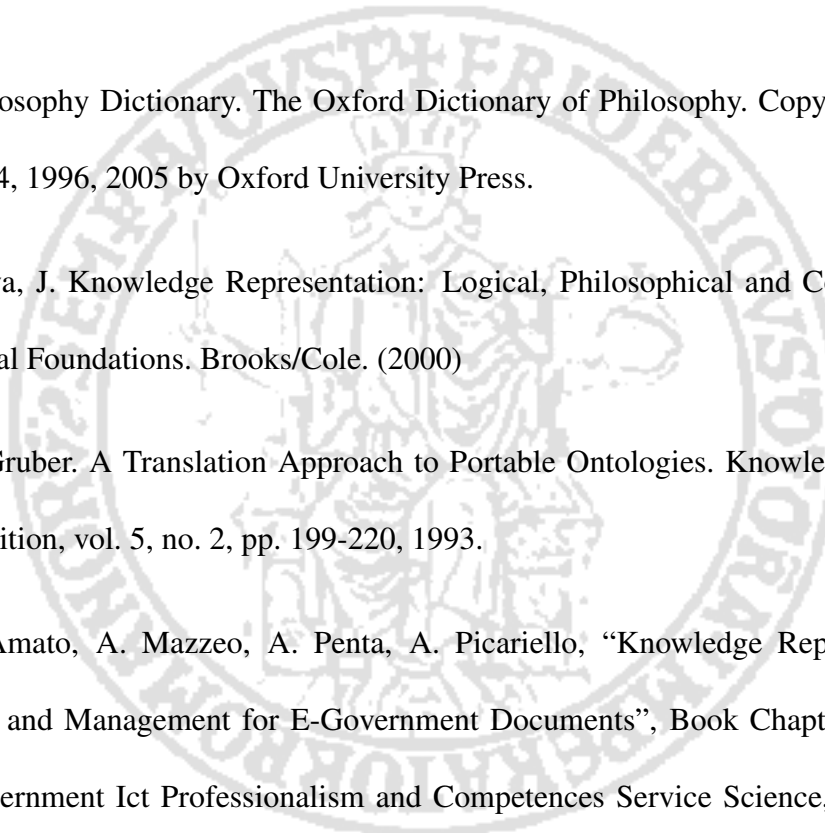
Conclusions

In this dissertation we have described an e-Government system based on a novel multimedia document model.

The proposed documents model, represented in RDF schema, is appropriate for the retrieval operations in different domains. The system is designed for the management of document belonging to specialized domain. The restricted area of specialization reduces the intrinsic semantic ambiguity of the words, related at the generalist domain, allowing more accurate information extraction operations.

We have implemented a prototypal version of the system that realize the described Information Retrieval and Presentation tasks for Long Term Preservation aims.

Bibliography

- 
- [1] Philosophy Dictionary. The Oxford Dictionary of Philosophy. Copyright © 1994, 1996, 2005 by Oxford University Press.
- [2] Sowa, J. Knowledge Representation: Logical, Philosophical and Computational Foundations. Brooks/Cole. (2000)
- [3] T. Gruber. A Translation Approach to Portable Ontologies. Knowledge Acquisition, vol. 5, no. 2, pp. 199-220, 1993.
- [4] F. Amato, A. Mazzeo, A. Penta, A. Picariello, “Knowledge Representation and Management for E-Government Documents”, Book Chapter of E-Government Ict Professionalism and Competences Service Science, pp.31–40, Springer Boston, 2008.
- [5] F. Amato, A. Mazzeo, V. Moscato, A. Picariello, “Information extraction from multimedia documents for E-Government” Book Chapter of Information Systems: People, Organizations, Institutions, and Technologies. Physica-Verlag. Springer. pp. 101-108. 2010. ISBN 978–3–7908–2149–9.

- [6] F. Amato, A. Mazzeo, A. Penta, A. Picariello, "A semantic document management system for legal applications", *International Journal of Web and Grid Services*, Vol. 4, No. 3, Inderscience Publishers, pp. 251–266(16), 2008.
- [7] R. Klischewski, M. Jeenicke, "Semantic Web technologies for information management within e-Government services", *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, vol., no., pp. 10 , 5-8 Jan. 2004.
- [8] L. Stojanovic, A. Abecker, N. Stojanovic, R. Studer, "On Managing Changes in the Ontology-Based E-Government" in *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE. 2004*, pp 1080-1097.
- [9] Deliberation of 13 dicembre 2001, n. 42, published on *Gazzetta Ufficiale della Repubblica Italiana* n. 296 of 21 dicembre 2001
- [10] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years", in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, 2000, pp. 1349- 1379.
- [11] M. S. Lew, N. Sebe, D. Djeraba, and J. Rain, "Content-based multimedia information retrieval: State of the art and challenges", *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, n.1, 2006 pp. 1–19.

- [12] R. Datta, and D. W. J. Joshi, "Image retrieval: ideas, influence, and trends of the new age", *ACM Computing Survey*, vol. 40, n. 2, pp. 5–64, 2008.
- [13] J. Z. Wang, J. Li, and G. Wiederhold, "Simplicity: Semantics sensitive integrated matching for pictures libraries. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, n. 1, pp. 1–16, 2001.
- [14] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blob world: image segmentation using expectation-maximization and its application to image querying", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, Issue 8, pp. 1026–1038, 2002.
- [15] J. S. Hare, and P. H. Lewis, "On image retrieval using salient regions with vector-spaces and latent semantics", *Image and Video Retrieval (CIVR 2005)*, Singapore, Springer Ed., 2005.
- [16] Elaine A. Rich, "Automata, computability and complexity: theory and applications", Prentice Hall, 2008, ISBN 0132288060, 9780132288064, pp. 1099.
- [17] M. Uschold , M. Gruninger "Ontologies and semantics for seamless connectivity" *ACM SIGMOD*, Vol.e 33 , I. 4 (December 2004), pp. 58 - 64, ISSN:0163-5808, Publisher ACM New York, NY, USA.
- [18] F. Amato, A. Mazzeo, V.Moscato, A. Picariello. "A System for Semantic Retrieval and Long Term Preservation of Multimedia Documents in E-

- Government Domain”. To Appear in International Journal of Web and Grid Services, Vol. 5, No. 4, Inderscience Publishers, pp. 323-338(16), 2009.
- [19] G. Boccignone, A. Chianese, V. Moscato, and A. Picariello. “Context-sensitive queries for image retrieval in digital libraries.”, Journal of Intelligent Information Systems, vol. 31, Issue 1, pp. 53–84, 2008.
- [20] Boccignone, G., Albanese, M., Moscato, V., and Picariello., A.: Image Similarity Based on Animate Vision: Information Path Matching. Multimedia Information Systems 2002: 66-75
- [21] P. Capasso, A. Chianese, V. Moscato, A. Penta, A. Picariello, “Automatic Categorization of Image Databases using Web Folksonomies”, in Proceedings of IEEE International Symposium on Multimedia, Dicembre 15-17, Berkley (California, USA), 2008.
- [22] J.M. Corridoni, A. Del Bimbo, A., and P. Pala, “Image retrieval by color semantics”, Multimedia Systems, vol. 7, n. 3, pp. 175–183, 1999.
- [23] C. Colombo, and A. Del Bimbo, “Visible image retrieval”, In L. Bergman and V. Castelli, eds., Image Databases, Search and Retrieval of Digital Imagery, Chapter 2, pp. 11-33, Wiley 2002.
- [24] S.Santini, “Evaluation Vademecum for Visual Information Systems,” Proc. of SPIE, vol. 3972, San Jose, USA, 2000

- [25] B. S. Manjunath and et al. Cortina, "Searching a 10 million images database", Technical report, Sep 2007.
- [26] J. Li, and J. Z. Wang, "Real-time computerized annotation of pictures", In Proc. ACM Int. Conf. on Multimedia, pp. 911-920, 2006.
- [27] Breuker, J. "A functional ontology of law", Journal of Artificial Intelligence and Law, Vol. 7, pp.341–361. 1994
- [28] P.R.S. Visser, T.J.M. Bench-Capon, "The formal specification of a legal ontology", in Proceedings of JURIX-96, Tilburg University Press, pp.15–24. 1996.
- [29] L.T. McCarty, "A language for legal discourse i. basic features", ICAIL '89: Proceedings of the 2nd International Conference on Artificial Intelligence and Law, New York, NY: ACM, pp.180–189. 1989.
- [30] Stamper, R. "The role of semantics in legal expert systems and legal reasoning", Ratio Juris, Vol. 4, No. 2, pp.219–244. 1991.
- [31] Tiscornia, D. "Some ontological tools to support legal regulatory compliance, with a case study", in Workshop on Regulatory Ontologies and the Modeling of Complaint Regulations (WORM CoRe), Springer LNCS, November 2003.

- [32] E. Riloff and R. Jones. “Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping”. In Proceedings of the Sixteenth National Conference on Artificial Intelligence. 1999
- [33] Jacobs, P. and Rau, L. (1990) “Scisor: extracting information from on-line news”, Comm ACM, Vol. 33, No. 11, pp.88–97.
- [34] J. R. Hobbs, D. Appelt, M. Tyson, J. Bear, and D. Israel, “Sri international: description of the Fastus system used for muc-4”, Fourth Message Understanding Conference, Morgan Kaufmann, pp.143–147. 1992.
- [35] S. Miike, E. Itoh, K. Ono, and K. Sumita, “A full-text retrieval system with a dynamic abstract generation function”, Proceedings SIGIR 94, pp.152–161. 1994
- [36] F. Amato, A. Mazzeo, A. Penta, A. Picariello, “Using NLP and Ontologies for Notary Document Management Systems”, in Proceedings of 19th International Conference on Database and Expert Systems Application, (DEXA), pp.67-71, 2008.
- [37] O. Udrea, V. S. Subrahmanian, Z. Majkic, “Probabilistic RDF” , in Proceedings of the 2006 IEEE International Conference on Information Reuse and Integration, IRI, pp. 172-177. 2006.

- [38] W. I. Grosky, "Managing Multimedia Information in Database Systems", Comm. Of ACM, vol. 40, n.12. 1997
- [39] L. Reeve, H.Han, Survey of semantic annotation platforms, in Proceedings of the ACM symposium on Applied computing, pp. 1634-1638. 2005.
- [40] A. Penta, A. Picariello, L. Tanca, "Towards a definition of an Image Ontology", in Proceedings of DEXA Workshops, pp. 74-78.2007.
- [41] A Bolasco S., della Ratta-Rinaldi F. (2004). Experiments on semantic categorisation of texts : analysis of positive and negative dimension. In Purnelle G., Fairon C., Dister A., Le poids des mots, Actes des 7es journées Internationales d'Analyse Statistique des Données Textuelles, UCL, Presses Universitaires de Louvain : 202-210
- [42] Park Y., Byrd R.J., Boguraev B.K., 2003, Towards Ontologies On Demand, IBM T.J. Watson Research Center 19 Skyline Dr., Hawthorne, NY 10562
- [43] Yukio Ohsawa, Nels E. Benson, Masahiko Yachida, "KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor", Advances in Digital Libraries Conference, IEEE, pp. 12, Fifth International Forum on Research and Technology Advances in Digital Libraries (ADL '98), 1998.

- [44] H. Cunningham and D. Maynard and V. Tablan. JAPE: a Java Annotation Patterns Engine (Second Edition). Technical report CS-00-10, University of Sheffield, Department of Computer Science, 2000.
- [45] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, M. Goranov. KIM – Semantic Annotation Platform. Book Chapter of The SemanticWeb - ISWC (2003). pp. 834 – 849. ISBN 978-3-540-20362-9-. Springer Berlin / Heidelberg.
- [46] , D. Tiscornia. Some ontological tools to support legal regulatory compliance, with a case study. Workshop on Regulatory Ontologies and the Modeling of Complaint Regulations (WORM CoRe 2003). Springer LNCS. November 2003.