Cultural and historical digital libraries dynamically mined from news archives

State of the Art

Project Reference No.	**FP7-215874**
Deliverable No.	**D2.1**
Workpackage no:	**WP2: User and System Requirements**
Nature:	**R (Report)**
Dissemination Level:	**PU (Public)**
Document version:	**05**
Date:	**05/05/2008**
Editor(s):	**Akrivi Katifori (NKUA/i)**
Document description:	This document summarizes the state of the art in the research areas addressed by Papyrus, the project research aims for progress beyond the state of the art along with the indicators for measuring this progress.

History

Version	Date	Reason	Revised by
01	2008-03-14	Initial Creation	Akrivi Katifori
02	2008-04-03	Section 4 updated with relevant SoA	Krishna Chandramouli
03	2008-04-20	Integration of individual contributions	Akrivi Katifori
04	2008-05-02	Feedback from AFP and DW integrated	Akrivi Katifori
05	2008-05-05	Revised Version after internal review	Akrivi Katifori

Authors List

Organisation	Name
NKUA/i	Akrivi Katifori (vivi@di.uoa.gr)
NKUA/h	Eirini Mergoupi-Savaidou
UNITN	John Mylopoulos(jm@cs.toronto.edu)
UNITN	Yannis Velegrakis (velgias@dit.unitn.it)
UNITN	Andrea Pressa
UNITN	Siarhei Bykau (bykau@disi.unitn.it)
QMUL	Krishna Chandramouli (krishna.chandramouli@elec.qmul.ac.uk)
QMUL	Ebroul Izquierdo (ebroul.izquierdo@elec.qmul.ac.uk)
CINECA	Roberta Turra (r.turra@cineca.it)
CINECA	Giorgio Pedrazzi (g.pedrazzi@cineca.it)
CINECA	Federico Giacanelli (f.giacanelli@cineca.it)
UNITN	Nadzeya Kiyavitskaya (nadzeya@disi.unitn.it)
UNITN	Luisa Mich (luisa.mich@economia.unitn.it)
AFP	Laurent Le Meur (Laurent.LEMEUR@afp.com)
DW	Martin Maass (Martin.Maass@dw-world.de)
ATC	George Kountourakis (g.kountourakis@atc.gr)

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Terms

CBIR	Content Based Image Retrieval
DOLCE	Descriptive Ontology for Linguistic and Cognitive Engineering
FC	Frequency Centroid
HZCRR	High Zero Crossing Rate Ratio
IPTC	International Press Telecommunications Council
KAA	Knowledge Assisted Analysis
MFCC	Mel Frequency Cepstral Coefficients
NITF	News Industry Text Format
NLI	Natural Language Inteface
NLP	Natural Language Processing
OWL	Web Ontology Language
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RF	Relevance Feedback
SCR	Silence Crossing Rate
SVM	Support Vector Machine
XML	Extensible Markup Language
ZCR	Zero Cross Rate

Executive Summary

This document presents the state of the art on the domains related with Papyrus. It is organized in five main parts, apart from the introduction and the conclusions.

The fist part offers a brief overview of existing web tools for historians to conduct research. Two main tool categories are distinguished, digital libraries and archives containing various forms of primary source material and websites dedicated to specific historical subjects, in some cases also offering access to some primary source material.

The second part presents research results and tools related to ontology issues including existing ontology models, ontology languages, storage modules, editors, versioning and evolution issues, visualization and matching and mapping between different ontologies.

Then a presentation of the state of the art on tools and methods for context analysis is being made, focusing on MPEG-7 visual features, content structuring tools, knowledge extraction, multi-modal analysis and text annotation

A brief presentation on semantic querying and browsing is made in the next section, covering formal queries and form-based, keyword and natural language interfaces.

Last but not least, a number of projects with a focus related to Papyrus are presented and their results are commented as to their suitability for re-use in the context of this project.

1. Introduction

Papyrus intends to be a dynamic digital library which will understand user queries in the context of a specific discipline, look for content in a domain alien to that discipline and return the results presented in a way useful and comprehensive to the user. To be able to achieve this, the source content has to be 'understood', which means analysed and modelled according to a domain ontology. The user query also has to be 'understood' and analysed following a model of this different discipline. Correspondences will then have to be found between the model of the source content and the realm of the user knowledge. Finally, the results have to be presented to the users in a useful and comprehensive manner according to their own 'model of understanding'.

Papyrus intends to showcase this approach with a specific pair of disciplines which can be illustrated as an apparent need and may prove to be an immediate exploitation opportunity even on its own. This proposed use case is the recovery of history from news digital content. To address these challenges Papyrus has planned an agenda of research in the following key areas:

- **Semantic multimedia analysis,** for the understanding of source content
- **Query processing**, for the understanding of the users demands
- **Knowledge mapping**, for corresponding concepts between the sources and the users
- **Presentation techniques**, for delivering the results in a comprehensive manner

In particular, knowledge technologies will have to be utilised for bridging the semantic gap between cultural heritage collections and their historical attributes, as expressed in news archives. Ontologies have been advocated as a means of semantic interoperability support between distributed applications and services by providing formal conceptualizations for specific domains. Together with other tools that have been developed in the context of the Semantic Web such as the RDF and OWL semantic mark-up languages and other knowledge representation and inference rule techniques, ontologies are expected to play a major role in deriving the appropriate level of new knowledge from what already exists in news archives.

Significant work will also be needed in the area of information extraction, both from the textual transcriptions of news but also from accompanying multimedia material (image and video) that is usually attached to many news items. One particularity of news contents residing in the archives of news agencies and public broadcasters is that some annotations have already been attached, usually in a manual fashion. This existing metadata could be exploited and used as a guide by special 'targeted' multimedia analysis methods to achieve realistic classification results.

This document presents an overview of the state of the art in all the aforementioned research areas. It provides a comprehensive view of current practices in both existing tools, employed by users and key research tendencies and results to be exploited by Papyrus. The document is organized as follows: Section 2 offers a brief overview of existing Web tools used by historians to conduct research.

Section 3 presents research results and tools related to ontology issues including existing ontology models, ontology languages, storage modules, editors, versioning and evolution issues, visualization and matching and mapping between different ontologies. Then a presentation of the state of the art on tools and methods for context analysis is being made, in Section 4, focusing on MPEG-7 visual features, content structuring tools, knowledge extraction, multi-modal analysis and text annotation A brief presentation on semantic querying and browsing is made in the following section, section 5, covering formal queries and form-based, keyword and natural language interfaces. Lastly, a number of projects with a focus related to Papyrus are presented and their results are commented as to their suitability for re-use in the context of this project.

2. Digital Tools for Historians

Digital libraries, and especially those that contain Historical Archive material are becoming a rather common and useful tool for the historians, since they offer easy access to primary source[1] material that is either difficult to find or fragile to use. Some of these digital libraries and archives bring together collections from various academic and cultural institutions.

Traditional historical research through printed or manuscript primary or secondary source material is being complimented by the use of digitized material in most cases even available through the WWW.

This section briefly presents some examples of digital tools currently used by historians to access material useful to their research.

2.1. Digital Archives

The recent great digitization effort has resulted in the creation of numerous Digital Libraries that may be accessible by historians.

The **European Culture Heritage Online**[2] (ECHO) is a collection of digital libraries of 50 scientific and cultural institutions worldwide, which contribute cultural heritage content as well as scholarly metadata. Access to digitized material, the majority of which concerns philosophy and science, is possible either by inserting key words or by browsing in thematic categories.

Internet Archive[3] is an internet library that offers permanent access for researchers, historians, and scholars interested in historical collections of digitized primary material in the form of text, audio and video. Search is based on keywords, corresponding to fixed thematic categories for each form of material. Internet Archive brings together digitized texts from the following initiatives: American Libraries, Canadian Libraries, Open Source Books, Gutenberg Project, Biodiversity Heritage Library, Children's Library and Additional Collections. Audio and video material comes from various collections. Internet Archive also offers free software and educational lectures in connection to certain academic institutions.

The **Perseus Digital Library**[4] of Tufts University contains primary material and secondary sources for research in the humanities, which are accessible by browsing or by inserting keywords in simple or advanced search. The digital library of Perseus includes various collections, such as the Classics Collection, the Renaissance Collection, the Bolles Collection, the California Collection, the Upper Midwest Collection, the Tufts History and the Boyle's Papers. The site also offers historical information on the related areas. A lot of the primary material is also available in text fromat. For the material in Greek, a transliterated version with comments was chosen, based on various bibliographic sources.

There are other many academic and cultural institutions that have started to digitize primary material, making it available in the form of PDF images. Greek digital libraries such as **Pergamos**[5] and

---

[1] **Primary source** is an original source of information that has been created by an authoritative individual with direct knowledge of the fact he/ she describes or impresses. On the contrary, **secondary source** discusses information originally presented elsewhere (primary sources or other secondary sources). For example, an article in the history of stem cells in the journal *History of Biology* is a secondary source for a historian who researches the history of stem cells by using as primary material references on stem cells found in news agencies items.

[2] http://echo.mpiwg-berlin.mpg.de/home

[3] http://www.archive.org/index.php

[4] http://www.perseus.tufts.edu/

[5] http://pergamos.lib.uoa.gr/dl/index

Hellinomnimon[6]**,** both of the National and Kapodistrian University of Athens, **Anemi**[7], of the University of Crete, **KENEF**[8], of the University of Ioannina, and others, offer simple and quick access to rich collections of digitized material of relevance to Modern Greek Studies scholars. Most of this material is old and rare. The researcher can browse the digital representations of old prints, manuscripts and visual material for historical, biographic and bibliographic information.

The **National Library of Greece** has developed a digital library[9] that is accessible through the Web. This digital library contains five Greek newspapers in digitized form, covering the period from the end of 19th to the middle of the 20th century. The material appears in PDF form. Each page of the newspapers corresponds to one PDF document. Newspaper issues are accessible either by browsing a calendar for each newspaper or by inserting the desirable date as a keyword. In addition to offering access to the digitized material, this digital library offers the researcher a useful real-time OCR tool for searching into the digitized material. To refine his/her search, the researcher can introduce one or more keywords (or word complexes), and count on the display the relevant pages of each newspaper. It is very helpful for the researcher that the keywords and phrases are underlined in the returned results, thereby avoiding the waste of time in order to read the whole page for the purpose of locating the word(s) he/she has asked for. This OCR tool is really useful for searching into the chaotic material of newspapers. Its major disadvantages are: a) that it does not return all the relevant results, and, b) that the interface of the display of the results is rather inconvenient for the user, since he/she has to open all the relevant pages in order to find what he/she is looking for.

The **BAM-Portal**[10] intends to provide an internet-based access point to cultural information in Germany. The partners of the BAM-Project started developing procedures to combine metadata from digital library catalogues, archival finding aids and museum inventories which allowed a simultaneous research. Academic persons and interested citizens can use the BAM-Portal as a first access point to the holdings of libraries, archives, museums and other kinds of institutions. They can see search results from these different types of institutions simultaneously. The results in the BAM-Portal are linked to the results in the original catalogues. Users may get more detailed information and are able to see the search result in context as well as available digitized documents.

Apart from access points to digitized archives and portals in national level, there exist also efforts to create unified portals in an international level that provide access to individual archives and museums.

MICHAEL[11] is a unified European portal for search in digital cultural heritage on a collection level. It draws on information from national portals, of which there are three online since 2006 (Italy, France, Britain), with Germany going online shortly and 16 more European countries within the year. Search is possible in collaborating collections that have to conform to certain standards. MICHAEL aims to allow comprehensive access to libraries, archives, museums and audiovisual content. It is EU commission funded. **Europeana**[12] (also commission funded) is a project that will produce a prototype for a European Digital Library. It will be going online in November 2008 and will provide access to roughly the same content as MICHAEL, but on an object level. There are many efforts ongoing in Europe and elsewhere to establish national Digital Libraries of comparable scope and approach, some of which will serve as national portals for the Europeana when available (e.g. the Deutsche Digitale Bibliothek DDB).

---

[6] http://www.lib.uoa.gr/hellinomnimon/

[7] http://anemi.lib.uoc.gr/

[8] http://www.kenef.phil.uoi.gr/en/index.php

[9] http://www/nlg.gr

[10] http://www.bam-portal.de/

[11] http://www.michael-culture.org/

[12] http://www.europeana.eu/

2.2. History Websites

Digital libraries contianing archive material are often part of general historical websites, which offer historical information about specific domains. This information may refer to particular eras, events, persons, etc, and may in some cases be connected to the primary source material.

The **VictorianWeb**[13] is a historical website for those who are interested in Victorian studies. It consists of both primary and secondary material, separated in thematic categories. In addition to original texts and visual material on Victorian literature, art, natural sciences and technology, the researcher may also find historical and bibliographic information, as well as historiographical essays on various topics of relevance to Victorian culture. The primary material does not appear in its original form, but it has been transformed in text format. Primary and secondary material is accessible by browsing a tree-structure of thematic categories or by inserting key words, yet the search tool is not very obvious in the site's interface.

Digital History[14] offers a source for interactive, multimedia history of the United States, from the American Revolution to the present. It makes available a variety of selective digitized sources in text, audio and video form, concerning American history and culture. The primary material is separated from the secondary material. The majority of primary sources have been transcribed in text form. Transcription in text was also chosen for newspapers' articles on more than 300 selected topics, although a PDF image for each of them is also available, but in low analysis. This PDF image may not enable the user to read the text from the original, but it provides a visual impression of the material under study. The site offers a search tool that is based on inserting key words for specific categories, but its interface directs the user to browsing.

Virginia Center for Digital History[15] is home to a number of digital projects, covering the whole range of American history. It offers access to primary material such as texts, audio and video, as well as to secondary material. It seeks to address the needs of educators, scholars, college students, and the general public. The material is accessible through browsing, in connection to specific topics of American history. No search tool is offered. The site also provides various links to the related topics.

Although the need for tools to relate secondary to primary source materials has been recognized by history researchers using electronic tools in their research, the aforementioned tools do not seem to be fulfil the basic requirements they seem to have for such a tool. These include a simple yet effective keyword search tool, the availability of the digitized material both in text and in its original form (as a scanned image for example), the existence of appropriate semantic annotation of the material and an appropriate way to represent the results.

In the context of Papyrus, historians basic requirements will be recorded and a set of desired functionalities will be compiled, in order to define the exact mechanisms and presentation methods that will provide access both to the ontologies and the primary source material itself.

[13] http://www.victorianweb.org/

[14] http://www.digitalhistory.uh.edu/

[15] http://www.vcdh.virginia.edu

3. Ontologies

According to [86], an **ontology** is an explicit specification of a conceptualization. The term "conceptualization" is defined as an abstract, simplified view of the world that needs to be represented for some purpose. It contains the objects, concepts and other entities that are presumed to exist in some area of interest and the relations that hold between them. The term "ontology" is borrowed from philosophy, where an ontology is a systematic account of Existence. For knowledge-based systems what "exists" is exactly that which can be (and has been) represented.

Therefore, as defined in [87], an ontology is a formal explicit description of concepts, or classes in a domain of discourse. Classes are organized in a hierarchical manner, expressing the inheritance relations between them. These inheritance relations form a hierarchy (or taxonomy) of classes. Properties -or slots- of each class describe various features and attributes of the class, and restrictions on slots (called facets or role descriptions) state conditions that must always hold to guarantee the semantic integrity of the ontology. Each slot has a type and could have a restricted number of allowed values. Allowed classes for slots of type Instance are often called a range of a slot.

This section presents in brief the state of the art in ontology related issues that are directly relevant to Papyrus. These include:

- Existing ontology models
- Ontology languages
- Storage and editing
- Versioning and Evolution
- Matching and Mapping
- Visualization

3.1. Existing ontologies

The main objective of Papyrus is to bridge the semantic gap between two different domains, history and news, through matching and mapping of their corresponding ontologies. To this end, in order to develop the appropriate ontology models for the two domains, existing general ontologies, as well as domain-specific ones will be taken into account.

3.1.1. General ontologies

Ontologies formalize a shared vocabulary about a domain [66]. Their importance stems from the fact that they offer well thought out terminologies for different domains that can be shared and reused. Tools for building and using ontologies are by now in abundance, based on a range of ontology languages with respect to expressiveness. A class of ontologies of special interest are the Foundational Ontologies, which are axiomatic theories that address very general domains [69]. The Descriptive Ontology for Linguistic and Cognitive Engineering (**DOLCE**[16]) is perhaps the best-known example among these. Ontologies can be classified into three main categories: upper, core, and domain:

- **Upper ontologies** (e.g., Cyc[17] and WordNet [68]) include general, domain-independent terms.
- **Core -- or intermediate – ontologies** cover broad domains, such as audiovisual phenomena.

[16] http://www.loa-cnr.it/DOLCE.html

[17] http://www.cyc.com/cyc/technology/whatiscyc

- **Domain ontologies**, on the other hand, are specific to a domain, such as manufacturing, history, or soccer).

A **folksonomy** is a user-generated taxonomy used to categorize and retrieve Web content such as Web pages, photographs and Web links, using open-ended labels called tags. Typically, folksonomies are Internet-based, but their use may occur in other contexts. Two widely cited examples of websites using folksonomic tagging are Flickr[18] and del.icio.us[19].

The **DS-MIRF** framework [70] proposes an OWL Upper Ontology that fully captures MPEG-7 MDS semantics. This ontology has been used for audiovisual content segmentation, resulting in the production of structured metadata that describe audiovisual content. These metadata are in OWL/RDF format and can be transformed into both MPEG-7 and TV-Anytime-compliant metadata, thus providing interoperability with software compliant with these standards.

SUMO[20] (Suggested Upper Merged Ontology) is an upper ontology which contains a set of general classes (such as Entity, Text, Language, Agent, Object, Temporal, Measure, Process, etc) and some properties (such as possesses, material, etc.) suitable for use in a broad range of domains. It is owned by IEEE and is freely available under an IEEE license. SUMO is written in KIF; an OWL version of SUMO is freely available for download[21]. SUMO is extended by a set of mid-level ontologies like **MILO** (MId-Level Ontology) which contains concepts like Newspaper, Publication, Sport, Government, etc. The union of these ontologies contains more than 20,000 terms.

PROTON[22] (PROTo ONtology) [402] is a lightweight upper-level ontology developed in the context of the EU project SEKT (http://www.sekt-project.com) for use on the Semantic Web. It is implemented in OWL Lite, and contains about 250 classes and 100 properties intended to provide basic elements for semantic annotation, indexing and retrieval. Some interesting concepts included in this ontology are for example: Event, Person, Location, Organization, Object, Product or Topic.

OpenCyc[23] is an open source version of Cyc. It consists of a general knowledge base and a reasoning engine. The knowledge base is written in CycL and can be accessed from Java programs using and open API. Version 1.0 of OpenCyc will contain more than 5,000 concepts (generic like for example TimeInterval, Place, Agent, Event, etc. and more specific like BussinessEvent, Ambulance, etc.) and 50,000 assertions. A free OWL version of OpenCyc Knowledge Base can be downloaded [24]. An online browser is available[25].

EuroWordNet[403][404] is a multilingual database, developed in the context of European projects LE-2 4003 and LE-4 8328, containing information for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The EuroWordNet Top Level ontology is made 63 classes, based on existing linguistic classifications, and developed to ease the task of encoding the different language resources in a uniform and interoperable way. It contains useful concepts such as Building, Vehicle, Instrument, Place, Human, Software, etc.

Ontologies have been largely studied in many EC projects, such as Knowledge Web or Open Knowledge. The same applies to the use of Formal Concept Analysis (FCA) (see for instance [71]) in information organization (see, e.g., the DELOS Network of Excellence in Digital Libraries).

---

[18] http://www.flickr.com/

[19] http://del.icio.us/

[20] http://www.ontologyportal.org/

[21] http://reliant.teknowledge.com/DAML/SUMO.owl

[22] http://proton.semanticweb.org

[23] http://www.opencyc.org

[24] http://sourceforge.net/project/showfiles.php?group_id=27274

[25] http://www.cycfoundation.org/concepts

3.1.2. News Ontologies

In the area of News Ontologies there is work available, mainly in the form of concept hierarchies.

NewsML[26] has been designed by the IPTC[27] (International Press Telecommunications Council) to provide a media-independent, structural framework for multi-media news. The need for NewsML came from the need for better and more consistent ways to structure, describe, manage and associate news content of different media types along their life-cycle, with rapid expansion of the Internet being a strong driving force.

At the heart of NewsML is the concept of the news item which can contain various different media – text, photos, graphics, video - together with meta-information that enables the recipient to understand the relationship between components and understand the roles of each component.

Everything the recipient might need to know about the content of the news provided can be included in NewsML's structure. For example, NewsML enables publishers to provide the same text in different languages; a video clip in different formats; or different resolutions of the same photograph. NewsML's standardized metadata sets include globally unique identifiers and version numbers that make it easy to track the evolution of a NewsItem over time, publication status (publishable, embargoed, etc.), administrative or descriptive properties such as creator, date of creation, copyright notice, subject and abstract.

NewsML defines default controlled vocabularies to ease the interoperability of implementations but it does not dictate which vocabulary is used on a given metadata property (IPTC Subject News Codes, ISO country codes etc.). Multiple vocabularies can even be utilised within the same NewsItem.

NewsML does not impose any content structure. For text objects in a NewsItem, the IPTC's News Industry Text Format (NITF) is recommended, but XHTML may also be chosen by a content provider.

NewsML is flexible and extensible and uses standard Internet naming conventions for identifying news objects in a NewsItem. As such, content does not have to actually be embedded within a NewsItem; pointers can be inserted to content held on a publisher's web site instead. This means subscribers retrieve the data only when they need to and this makes NewsML bandwidth-efficient.

A new major version of this standard, named **NewsML-G2**, has been released in 2008. It is a member of a family of complementary IPTC news exchange format standards - collectively known as **G2-Standards** which also offers a standard representation of news events and another for sports results and statistics.

NewsML-G2 has been built around an object model expressed as UML graphs, which may be easily mapped to a formal ontology model expressed in OWL. This model is made of two parts: a structural model representing news items and news packages, and a basic model of concepts useful for the annotation of general news, e.g. people, organisations and locations.

The **IPTC Subject News Codes**[28] are sets of topics (aka topical subjects) to be assigned as metadata values to news objects like text, photographs, graphics, audio- and video assets. This allows for a consistent coding of news metadata over the course of time. This 3-levels taxonomy has currently 1,300 terms in it. Each term corresponds to a numeric code and is associated with labels in English, French, German and Spanish. The use of IPTC Subject News Codes is recommended by the IPTC for the classification of NewsML documents.

Within the MESH[29] project, an OWL ontology has been built, extending the IPTC taxonomy with terms and categories in two areas ("Disaster and accidents" and "Unrest conflicts and war").

---

[26] http://www.newsml.org

[27] http://www.iptc.org

[28] http://www.iptc.org/NewsCodes/index.php

[29] http://www.mesh-ip.eu/?Page=Project

Another news ontology has been developed by the UMBC ebiquity research group[30] and is available as an OWL News Ontology[31]. It does not seem however as particularly elaborated

The **NEWS Ontology** [64] has been developed in the context of the NEWS[32] project [65]. It covers the main concepts required in the news domain. It is a lightweight RDFS ontology and provides the basic classes, properties and instances for news item categorization and content annotation. Standards from the journalism world, like NITF, NewsML, have been used as sources in the initial knowledge capture process, and the generic SUMO ontology has been used as a base for the content annotation module of the ontology. Another point also considered is multilingualism: the concepts included in the NEWS ontology have associated labels and descriptions in several languages (more specifically, Spanish, Italian and English). These labels and descriptions are useful, for instance, in order to implement semantic search facilities, which require that users disambiguate their queries.

3.1.3. History Ontologies

As opposed to the News Domain, where there are basic News ontologies already available that could be used and extended in the context of Papyrus, in the History domain there is not a lot of work available. The closest model to a historical ontology could be considered the CIDOC-CRM.

The **CIDOC Conceptual Reference Model[33] (CRM)** provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. IT is intended to promote a shared understanding of cultural heritage information by providing a common and extensible semantic framework that any cultural heritage information can be mapped to. It is intended to be a common language for domain experts and implementers to formulate requirements for information systems and to serve as a guide for good practice of conceptual modelling. In this way, it can provide the "semantic glue" needed to mediate between different sources of cultural heritage information, such as that published by museums, libraries and archives.

However, this model is oriented to museum objects, whereas in this case Papyrus focuses on History itself. To this end, the main challenge in Papyrus is to create ontologies that satisfy the two main characteristics of an international History ontology, **multilingualism** and **temporal characteristics**. This need reflects a general issue for ontologies that may be useful in many domains related to digital libraries, not only the ones, History and News, that are used as an example by Papyrus. The content of most Digital Libraries has a temporal dimension, may cover large time-spans and include changes in terminology in the passage of time. Multilingual ontologies are of even greater necessity, especially in a large, multinational community as the European one. To this end, the research in the context of Papyrus on this two main ontology issues is of a more general interest than satisfying the needs of History through News.

On multilingual ontologies, work has been done in the support of label translation, such as the NeOn toolkit[34] and the work in [400] and [401]. However, the problem of multiple translations in Papyrus has a more complex perspective that needs to be investigated. Different concepts appear in different countries in different time periods and, in some cases, with slightly different semantics. To this end, the problem of multiple languages in Historical ontologies should be handled in accordance with the temporal aspect of modelling historical information through an ontology.

3.2. Ontology Languages

[30] http://ebiquity.umbc.edu/

[31] http://ebiquity.umbc.edu/ontology/news.owl

[32] http://www.news-project.com

[33] http://cidoc.ics.forth.gr/

[34] http://www.neon-toolkit.org/

The interchange of ontologies across the World Wide Web (WWW) and the cooperation among heterogeneous agents placed on it is the main reason for the development of a new set of ontology specification languages, based on Web standards. These languages aim to represent the knowledge contained in an ontology in a simple and human-readable way, as well as allow for the interchange of ontologies across the Web. An overview of several of these may be found in [53]. Here we present two of the main Web ontology languages, both widely used and accepted by the Semantic Web community, as well as supported by existing tools, RDF/RDFS and OWL.

RDF[35] **(Resource Description Framework)** is a set of W3C[36] specifications originally designed as a metadata data model, but which has come to be used as a general method of modeling information through a variety of syntax formats.

The RDF metadata model is based upon the idea of making statements about Web resources in the form of subject-predicate-object expressions, called triples in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object. RDF is an abstract model with several serialization formats (i.e., file formats), and so the particular way in which a resource or triple is encoded varies from format to format.

The following example presents a description in RDF of a specific music album.

```
<rdf:Description
 rdf:about="http://www.recshop.fake/cd/Hide your heart">
        <cd:artist>Bonnie Tyler</cd:artist>
        <cd:country>UK</cd:country>
        <cd:company>CBS Records</cd:company>
        <cd:price>9.90</cd:price>
        <cd:year>1988</cd:year>
</rdf:Description>
```

RDF properties may be thought of as attributes of resources and may also represent relationships between resources. RDF however, provides no mechanisms for describing these properties, nor does it provide any mechanisms for describing the relationships between these properties and other resources. That is the role of the RDF vocabulary description language, RDF Schema.

RDF Schema defines classes and properties that may be used to describe classes, properties and other resources. It provides mechanisms for describing groups of related resources and the relationships between these resources. RDF Schema vocabulary descriptions are written in RDF using the terms described in this document. These resources are used to determine characteristics of other resources, such as the domains and ranges of properties.

OWL[37] was developed on top of RDF and borrowed from DAML+OIL[38]. Like RDF, OWL is the standard recommended by W3C for Semantic Web. OWL is powerful in expression, but complex for computation. To compromise between expressive power and acceptable computational complexity, OWL has three increasingly-expressive sublanguages: OWL Lite, OWL DL, and OWL Full. Among them, OWL Lite is a subset of OWL DL, and OWL DL is a subset of OWL Full.

[35] http://www.w3.org/TR/rdf-syntax-grammar/

[36] http://www.w3.org/

[37] http://www.w3.org/2004/OWL/

[38] http://www.daml.org/

- **OWL Lite** supports those users primarily needing a classification hierarchy and simple constraints. For example, while it supports cardinality constraints, it only permits cardinality values of 0 or 1. It should be simpler to provide tool support for OWL Lite than its more expressive relatives, and OWL Lite provides a quick migration path for thesauri and other taxonomies. Owl Lite also has a lower formal complexity than OWL DL.
- **OWL DL** supports those users who want the maximum expressiveness while retaining computational completeness (all conclusions are guaranteed to be computable) and decidability (all computations will finish in finite time). OWL DL includes all OWL language constructs, but they can be used only under certain restrictions (for example, while a class may be a subclass of many classes, a class cannot be an instance of another class). OWL DL is so named due to its correspondence with description logics, a field of research that has studied the logics that form the formal foundation of OWL.
- **OWL Full** is meant for users who want maximum expressiveness and the syntactic freedom of RDF with no computational guarantees. For example, in OWL Full a class can be treated simultaneously as a collection of individuals and as an individual in its own right. OWL Full allows an ontology to augment the meaning of the pre-defined (RDF or OWL) vocabulary. It is unlikely that any reasoning software will be able to support complete reasoning for every feature of OWL Full.

The next example presents the definition of two sub-classes in OWL.

```
<owl:Class rdf:ID="Lens">
    <rdfs:subClassOf rdf:resource="#PurchaseableItem"/>
</owl:Class>

<owl:Class rdf:ID="Camera">
    <rdfs:subClassOf rdf:resource="#PurchaseableItem"/>
</owl:Class>
```

The exact needs of Papyrus converning the appropriate language to be used for the History and News Ontologies will be defined after the user requirements will have been recorded. After the thorough investigation of the issue, the appropriate ontology language will be selected, in combination with an effective tool for storage. A selection of ontology storage tools is presented in the following section.

3.3. Ontology Storage

A very important aspect of ontology management is the storage and retrieval of ontologies. This issue requires development of ontology management systems, responsible for 'semantic'-based ontology storage. Efficient re-use of knowledge in a closed context requires a library system for ontologies. An ontology management system should take care of storage, identification, edition, and retrieval of ontologies used for similar applications. As the number of various ontologies is growing, maintaining and re-organizing these ontologies so as to facilitate the reuse of knowledge is challenging. A pre-requisite for the brake-through of ontology technology is the support of methods and tools that enable their effective and efficient development. A key aspect in enabling this is successful reuse of ontologies. Being developed for supporting knowledge sharing and reuse, it is the lack in proper support of ontology reuse that hampers a broader dissemination of the ontology. Ontology library systems are an important tool to group and re-organize ontologies for further reuse, integration, maintenance, mapping and versioning.

The main criteria for an ontology management system are the following:

- Supporting ontology reuse by open and central storage, identification and versioning.
- Supporting ontology reuse by providing smooth access to existing ontologies and by providing advanced support in adapting ontologies to certain domain and task specific circumstances (instead of forcing to develop such ontologies from scratch).
- Supporting ontology reuse by fully employing the power of standardization. Providing access to standardized upper-layer ontologies and representation languages is one of the main steps in bringing knowledge sharing and reuse to its full potential.

Most ontology storage systems have either a client/server-based architecture aiming for remote accessing and collaborative editing (WebOnto[39], OntoLingua[40], DAML Ontology Library[41], Protégé) or a Web accessible architecture (SHOE[42], IEEE SUO[43]). Ontology Server has a database-based architecture. Most of the ontologies are classified or indexed. They are stored in the modular structured library (or lattice of ontologies). WebOnto, OntoLingua and ONIONS all emphasize the importance of modular structure for ontology library system, which pave the road for the reuse of ontology, management of ontology, reorganization of ontology library system.

A detailed report on several of the available tools may be found in [51]. Here we present briefly three of the most prominent candidates for use in the context of Papyrus.

Sesame[44] [55] is an open source RDF framework with support for RDF Schema inferencing and querying. Originally, it was developed by Aduna[45] (then known as Aidministrator) as a research prototype for the EU research project On-To-Knowledge[46]. Now, it is further developed and maintained by Aduna in cooperation with NLnet Foundation[47], developers from Ontotext[48], and a number of volunteer developers who contribute ideas, bug reports and fixes.

---

[39] http://kmi.open.ac.uk/projects/webonto/

[40] http://www.ksl.stanford.edu/software/ontolingua/

[41] http://www.daml.org/ontologies/

[42] http://www.cs.umd.edu/projects/plus/**SHOE**/

[43] http://suo.ieee.org/

[44] http://www.openrdf.org/

[45] http://www.aduna-software.com

[46] http://www.ontoknowledge.org/

[47] http://www.nlnet.nl/

[48] http://www.ontotext.com/

Sesame has been designed with flexibility in mind. It can be deployed on top of a variety of storage systems (relational databases, in-memory, file systems, keyword indexers, etc.), and offers a large selection of tools to developers to leverage the power of RDF and RDF Schema, such as a flexible access API, which supports both local and remote (through HTTP or RMI) access, and several query languages, of which SeRQL is the most powerful one.

There is a Protégé plug-in[49] for Sesame, but not very actively developed.

Jena[50] [58][59] is a Java framework for building Semantic Web applications. It provides a programmatic environment for RDF, RDFS and OWL, SPARQL and includes a rule-based inference engine. Jena is open source and grown out of work with the HP Labs Semantic Web Programme. It includes:

- A RDF API
- Reading and writing RDF in RDF/XML, N3 and N-Triples
- An OWL API
- In-memory and persistent storage
- A SPARQL query engine

Jena2 [61] is the second generation of the Jena toolkit. It conforms to the revised RDF specification, has new capabilities and a new internal architecture. A design principle for Jena2 was to minimize changes to the API from Jena to Jena2. The Jena database subsystem implements persistence for RDF graphs using an SQL database through a JDBC connection.

Jena has also been integrated in Protégé-OWL, as explained in [63].

There are other commercial tools like the ones created by Oracle, Mondeca and Ontoprise. We present them here as an example of the tools available commersialy that will be taken into account.

Oracle[51]'s proposal stores RDF triples in the Oracle database as a logical network (using the Oracle Spatial Network Data Model). Oracle 10g supports directed and un-directed logical graphs (networks) as part of Oracle Spatial Network Data Model (NDM). The proposed RDF data model maps RDF triples to a logical network managed by NDM. In addition to the core data, a catalogue service is provided: by maintaining information about different RDF models, including the namespaces used in these models. RDF triple data is mapped onto a graph by storing subjects and objects as nodes, and properties as links. The storage for RDF data is managed by Oracle: all the RDF data is managed in a central schema, and user-level access functions and constructors are provided to query and update the RDF data. There is one universe for all RDF data stored in the database. Each RDF triple: {subject, property, object} is treated as one unique database object. As a result, a single RDF document comprising a number of triples will result in multiple database objects. Oracle also offers a sample Protégé plug-in[52] for its RDF store.

Mondeca[53] ITM e-Knowledge is a knowledge representation management application based on Semantic Web technology (Web 3.0) and ontologies. With a complete modelling capability according to the area of activity, the tool enables all types of concepts to be described and organised into classes, and all types of relationships between concepts to be managed, ensuring rapid deployment of an operational solution that is perfectly suited to the specialist domain. Designed for the web and for SOA (service-oriented architecture), the tool enables advanced information access and processing services to be offered to users, and also the construction of automated services using web services.

---

[49] http://protege.stanford.edu/plugins/rdfs-db/

[50] http://jena.sourceforge.net/

[51] www.oracle.com

[52] http://www.oracle.com/technology/tech/semantic_technologies/sample_code/index.html

[53] http://www.mondeca.com/index.php/en

OntoPrise OntoStudio[54] is a professional development environment for modelling ontologies and administrating ontology-based solutions that allows for the integration of multiple heterogeneous data sources. The product is based on a modular design and supports the development of defined modules as well as use-based customization. Eclipse, the open source editor framework, provides the implementation base for the product.

The use of the appropriate ontology storage module is essential in the context of Papyrus. The aforementioned options will be thoroughly investigated as to their effectiveness, compatibility with other selected tools, robustness and support. Ontology storage evaluations, like the one in [54] will be also taken into account.

3.4. Ontology Editors

As ontologies are becoming a prominent tool for knowledge representation, several editors have been developed to facilitate their creation and management. A survey of existing tools may be found in [56], featuring a comprehensive table[55] of existing editors features, with a revised, more updated version in [57], as well as the corresponding table[56].

In this section we will present in more detail two of the most prominent ones in the scientific community, both supporting the creation of third party plug-ins: Protégé and NeOn Toolkit.

Protégé[57] is a graphical and interactive ontology-design and knowledge-acquisition environment that is being developed by the Stanford Medical Informatics group (SMI) at Stanford University. Its knowledge model is compatible with OKBC [50]. Its component-based architecture enables system builders to add new functionality by creating appropriate plug-ins. It is referred in many surveys as the most complete ontology editor available at the moment. The current version is the 3.3, but there are a 3.4 in beta testing and a 4 in alpha, which development is performed by a community different from the original creators. The key point of the Protégé is the extensibility of the code. It is an open source project that can be downloaded free through the Web. Its provides an API that allows the creation of applications and plug-ins. The API is written in Java and ontology data can be saved in various formats, including relational database. This allowed the development of many plug-ins that are freely available for download. Originally Protégé was designed for the ontology language OKBC, but has lately been extended to provide OWL support in its various forms (lite, DL and full).

Due to its extensibility, the plug-ins that have been developed include multiple different features, including, but not limited to:

- Support for collaborative development
- Prompt tab enables mapping, merging and versioning capabilities when managing more than one ontologies
- Web browsable front-end for the Protégé basic functionalities
- Visualization tools, for ontologies in various formats
- Relational data importing functionalities
- Query tabs for retrieving information from the underlying knowledge base
- Java class and UML diagram generation
- Automated reasoning, such as support for JESS rules over OWL ontologies
- Import/export functionality between ontologies and XML schemas

[54] http://www.ontoprise.de/content/e1171/e1249/index_eng.html

[55] http://www.xml.com/2002/11/06/Ontology_Editor_Survey.html

[56] http://www.xml.com/2004/07/14/examples/Ontology_Editor_Survey_2004_Table_-_Michael_Denny.pdf

[57] http://protege.stanford.edu/

The Protégé-OWL (Figure 3-1) editor is an extension of Protégé that supports the Web Ontology Language (OWL). The Protégé-OWL editor enables users to:

- Load and save OWL and RDF ontologies.
- Edit and visualize classes, properties, and SWRL rules.
- Define logical class characteristics as OWL expressions.
- Execute reasoners such as description logic classifiers.
- Edit OWL individuals for Semantic Web markup.

Protégé-OWL's flexible architecture makes it easy to configure and extend the tool. Protégé-OWL is tightly integrated with Jena and has an open-source Java API for the development of custom-tailored user interface components or arbitrary Semantic Web services.

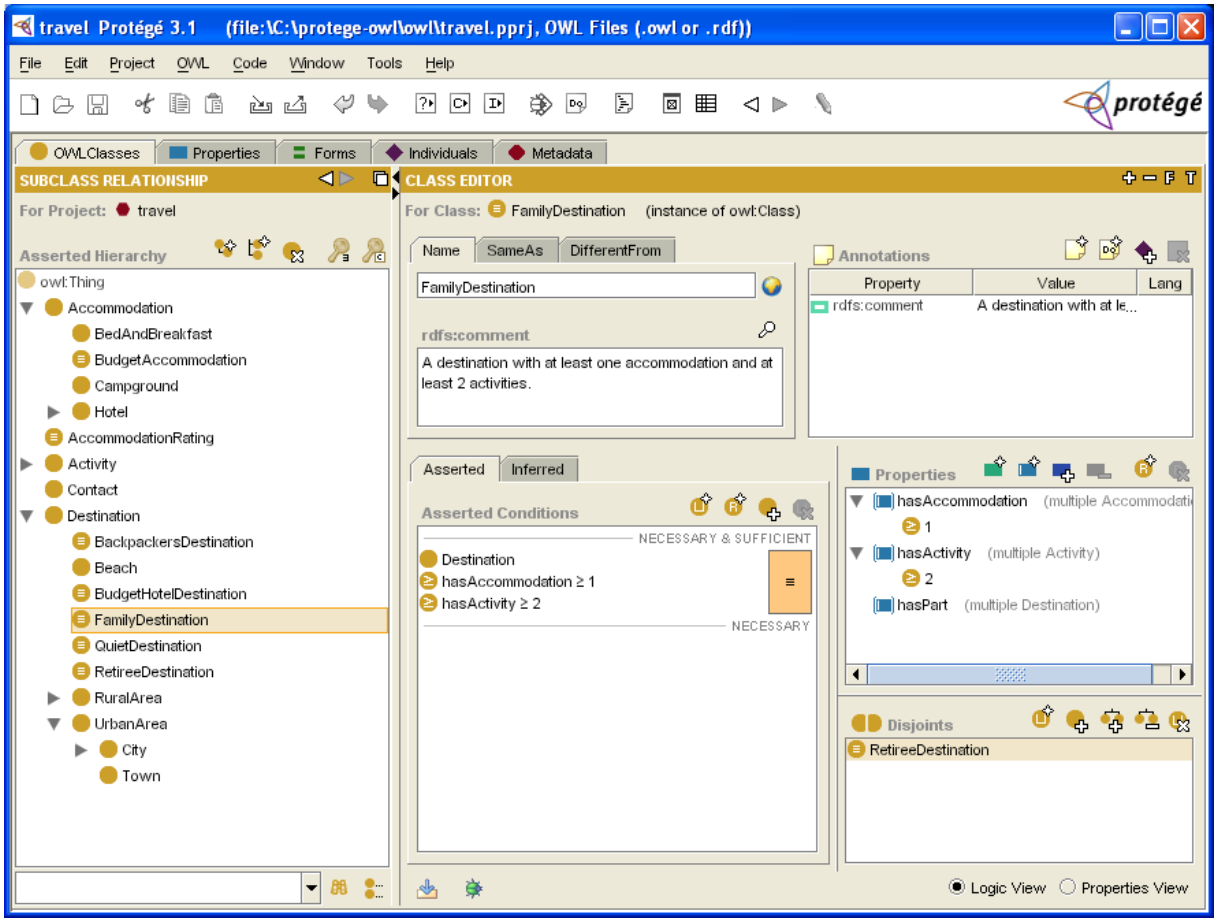

Figure 3-1. The Protégé OWL Editor

The **NeOn toolkit**[58] (Figure 3-2) is an open source extensible Ontology Engineering Environment. It has been developed in the context of the NeOn project[59], involving 14 European partners and co-funded by the European Commission's Sixth Framework Programme. NeOn started in March 2006 and has a duration of 4 years. Its aim is to advance the state of the art in using ontologies for large-scale semantic applications in the distributed organizations.

[58] http://www.neon-toolkit.org/

[59] http://www.neon-project.org/web-content/

NeOn toolkit is part of the reference implementation of the NeOn architecture. It contains plug-ins for ontology management and visualization. The core features include:

- Basic Editing: Editing Schema
- Visualization/Browsing
- Import/Export: F-Logic, (subsets of) RDF(S) and OWL

Neon also supports the creation of plug-ins.

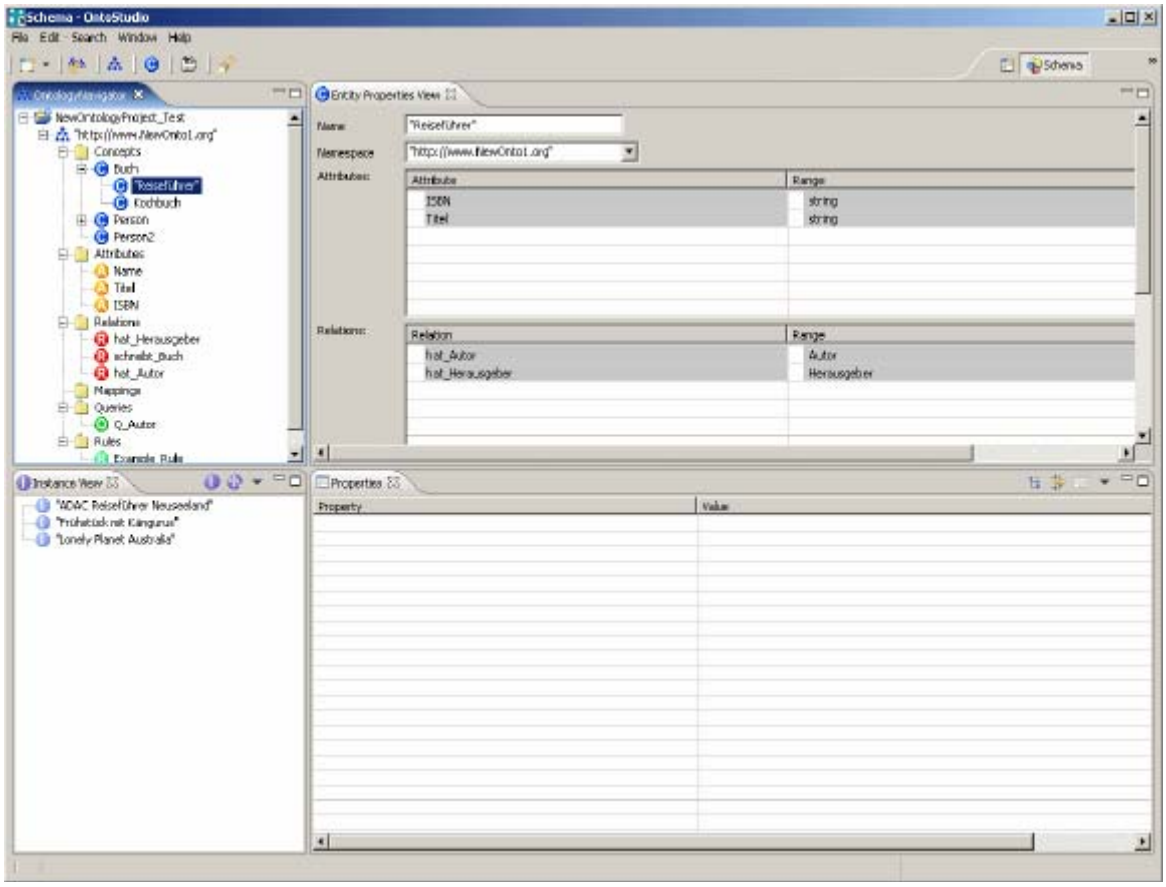

Figure 3-2. A view o the NeOn tooltik

Swoop is another editor, at the moment actively developed as an open-source project[60], that aims at providing rapid and easy development for OWL ontologies. It features a web browser look and feel, inline editing and features for specific OWL requirements.

[60] http://code.google.com/p/swoop/

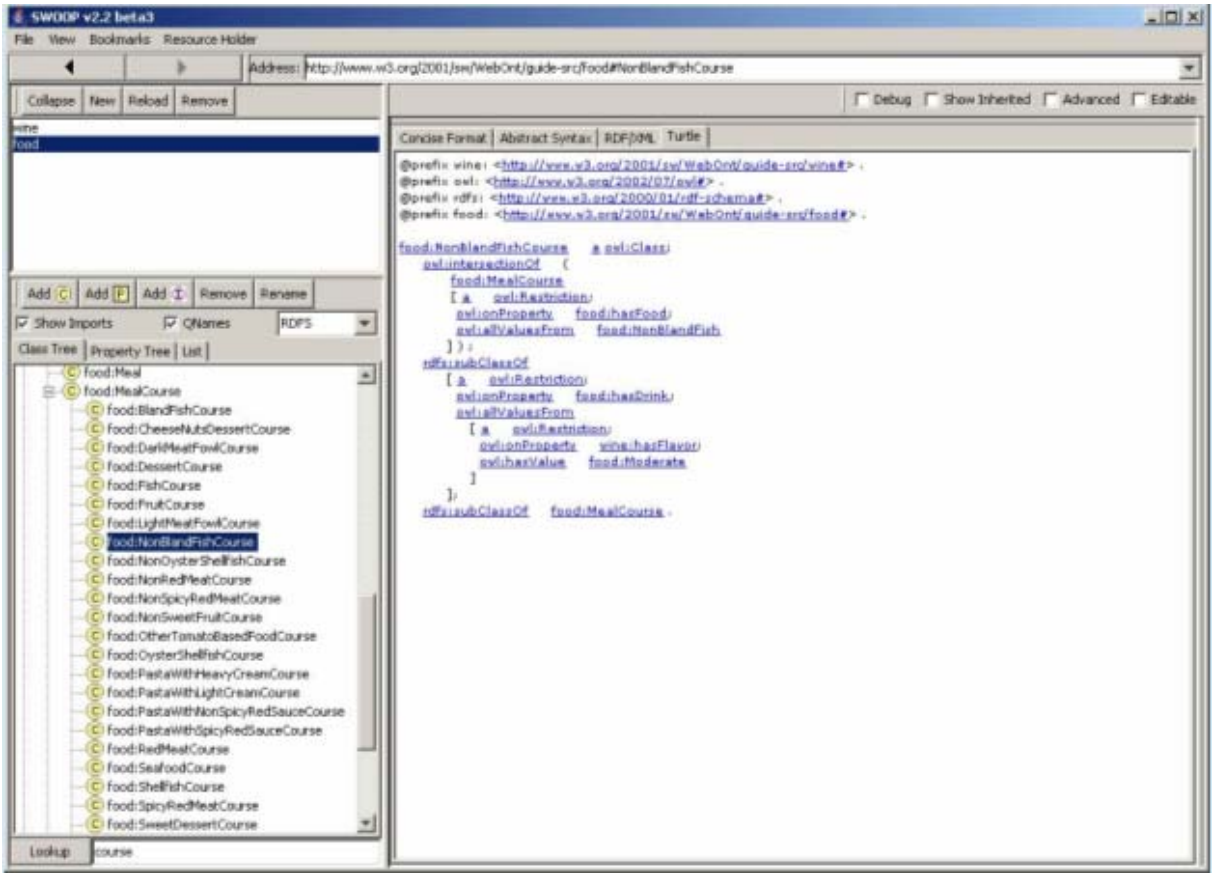

Figure 3-3. The SWOOP ontology editor

According to the user needs and requirements, all these tools will be taken into account for use and extension in the context of Papyrus.

3.5. Ontology Versioning and Evolution

In contrast with generic practice of maintaining *ontology snapshots*, i.e. only the most recent version of ontologies is maintained in the system, **ontology versioning** is accommodated when an ontology management system allows for handling of ontology changes by creating and managing different versions of it. **Ontology evolution** is accommodated when an ontology management system facilitates the modification of an ontology by preserving its consistency [32].

There are several works relative to the subjects of ontology versioning and evolution. A comparative evaluation of ontology editors concerning the subject of functions supporting ontology evolution may be found in [38]. Although Papyrus exact needs and requirements on these issues will be fully defined after the completion of the requirements and specifications phase, here we present some prominent examples of the work produced in this domain so far, in order to have a more clear idea on available tools and methods to be used if needed later in the project.

In [42] the authors discuss the problems associated with managing ontologies in distributed environments such as the Web. They present SHOE, a web-based knowledge representation language that supports multiple versions of ontologies. SHOE is described in the terms of a logic that separates data from ontologies and allows ontologies to provide different perspectives on the data. The paper presents the features of SHOE that address ontology versioning, the effects of ontology revision on SHOE web pages, and methods for implementing ontology integration using SHOE s extension and version mechanisms.

In [43] the authors discuss the problem of ontology versioning based on work done in database schema versioning and program interface versioning. They also propose building blocks for the two important aspects of a versioning mechanism: ontology identification and change specification.

[44] discusses OntoView, a web-based change management system for ontologies. OntoView provides a transparent interface to different versions of ontologies, by maintaining not only the transformations between them, but also the conceptual relation between concepts in different versions. It uses several rules to find changes in ontologies and it visualizes them—and some of their possible consequences—in the file representations. The user is able to specify the conceptual implication of the differences, which allows the interoperability of data that is described by the ontologies. The paper describes the system and presents the mechanism that we used to find and classify changes in RDFS / DAML ontologies. It also shows how users can specify the conceptual implication of changes to help interoperability.

In [39] the authors identified a possible evolution process. To represent changes, they introduced in [40] three levels of abstractions for ontology changes fro the KAON[61] language. They distinguish elementary changes (modifications to one single ontology entity), composite changes (modifications to the direct neighborhood of the ontology entity) and complex changes (modifications to an arbitrary set of ontology entities). To support the understanding of evolution, their framework provides an overview of all applied changes. In [41], the notion of a version log is introduced and a Change Definition Language (CDL) is presented for the OWL DL ontology language. The version log keeps track of the changes whereas the CDL allows the users to define the meaning of changes in a formal way.

The system **PromptDiff** [33] has been developed in the context of a collaborative environment for managing ontologies in order to support ontology versioning and is available as a Protégé[62] plug-in. Given two versions of an ontology, it allows users to: (1) examine the changes between versions visually; (2) understand the potential effects of changes on applications; and (3) accept or reject changes. The visualization of differences is based on the Microsoft Word Compare Documents paradigm. At the center of the system is a Change and Annotation Ontology (CHAO) with instances recording specific changes and meta-information about them. The two versions are presented the one

[61] http://kaon.semanticweb.org

[62] http://protege.stanford.edu

next to the other with highlighting on the parts where changes have occurred. **PromptViz** [45] is a tool providing advanced visualizations using treemaps to help users understand the location, impact, type and extent of changes that have occurred between versions on an ontology. COVE [34] is another collaborative environment for managing and merging ontologies, based on OWL.

In [35], a different approach for reconciling the different ontology versions with each other is introduced. The presented framework, that aims to provide means for reasoning based on a complete versioning history, includes the generic notion of *change bridge* for describing ontology resource chances, and a basic set of particular change bridge types that constitute the class hierarchy of a change bridge ontology. Ontology changes are represented as instances of the change types relating concepts in successive ontology versions with each other. The change bridge ontology is represented using RDF. In [37] the changes in an ontology are handled as database operations and triggers, whereas in [36] a formalization of operations that change the knowledge base is presented, in order to support the evolution of ontologies. An abstract data type knowledgebase is defined, which contains a description logic representation and a basic set of operations to work on it. These can be extended or changed to satisfy local needs. It also leads to a formal introduction of operations for knowledge engineering, which can be used for ontology life-cycle management.

Papyrus, as a tool attempting to reconcile to different domains through their ontologies, needs to consider to main issues related to versioning and evolution. Firstly, the history and news ontologies will be created by teams of experts. Thus, tools to support collaborative editing and basic versioning are essential. Preliminary interviews with historians so far have not concluded in the need for support for the logging and presentation of complex evolution information at the ontology design stage. However, the main theme of papyrus is History through News and, as a result, a different kind of evolution is essential, that of modelling entity (class and instance) evolution in the time period that the history ontology covers.

This issue has not been thoroughly investigated and there are few works available. An approach to modelling history knowledge through an ontology is presented in [22]. They introduce validity periods to entities as well as temporal data types to represent changing properties and relations between entities. In [46], [47] and [48] an ontological method for representing spatio-temporal changes in regions that essentially define a geo-ontology time series has been proposed. These approaches will be taken into account for the modelling, management and visualization of historical entities.

3.6. Ontology Matching and Mapping

To achieve interoperability one needs to be able to translate data or knowledge from one form of representation to another. This translation is typically described though a set of abstract assertions that are referred to as *mappings.* Although mapping is a generic term, it has been mostly used between database schemas. Generating mappings between two schemas (or ontologies) is a laborious and time-consuming process. Thus, there has been an effort to provide the data administrators with tools that facilitate that task.

A required first step in specifying mappings is a schema (or ontology in the case of ontologies) matching process. Schema matching is an old problem that has been studied in different contexts. A matching provides simple associations between concepts of two different schemas/ontologies. In schema matching, two schemas are compared and the result is a set of binary relationships between their elements. To perform such a matching, different techniques have been proposed in the literature. Most of them have considered matching between database schema elements and are based on some syntactic, lexical, or data instance information to realize the relationship between the different elements. Rahm and Bernstein provided a nice survey of the most common such techniques [79].

Melnik et al. [78] suggest a structural algorithm that can be used for matching schemas. The algorithm is based on the following idea. First, the schemas to be matched are converted into directed labeled graphs. These graphs are used in an iterative fix-point computation whose results indicate

what nodes in one graph are similar to nodes in the second graph. For computing the similarities, they rely on the intuition that elements of two distinct models are similar when their adjacent elements are similar. In other words, a part of the similarity of two elements propagates to their respective neighbors. The spreading of similarities in the matched models is reminiscent to the way Internet Protocol packets flood the network in broadcast communication. For this reason, Melnik et al. call their algorithm similarity flooding. With such a tool, matching is not done entirely automatically. Instead, the tool assists human developers in matching by suggesting plausible match candidates for the elements of a schema. Using a graphical interface, the user adjusts the proposed match result by removing or adding lines connecting the elements of two schemas.

The **Cupid** [77] and **COMA** [76] systems use a number of different approaches and combine their output in order to achieve better matching suggestion. They employ algorithms that use linguistic reasoning to match attributes based on their names or their place in a hierarchical structure. In addition, Cupid uses information about types, optionality, cardinalities, etc. It also tries to match schema constraints such as keys or referential constraints. He and Chang [75] noticed that, in specific communities, the schema vocabulary tends to converge at a relatively small size. For such cases, they propose the use of probabilistic techniques to infer matchings.

If some data is already stored in the target schema, one may also use a data mining technique that matches data values or their characteristics to find binary relationships between the schema elements [74]. The **LSD** system [73] uses a multi-level learning scheme to find 1:1 binary relationships between XML DTD tags. A number of base learners that use different instance-level matching schemes are trained to assign tags of a mediated XML schema to data instances of a source schema. Then, a meta-learner combines the predictions of the base learners. Techniques that are data-value based are important for cases where determining correspondences based on the names of the schema elements is difficult. Based on that observation, Kang and Naughton [72] measure the pair-wise attribute correlations in the schema elements to be matched and construct a dependency graph as a measure of the dependency between attributes. Then they find matching node pairs in the dependency graphs by running a graph matching algorithm.

Although schema matching is aided tremendously by the many recent and interesting results in the research community, full automation of this task has not been achieved. The reason is that even matches as plausible as Company to company can be deemed as incorrect by a data warehouse designer who knows that the first schema element is used to represent names of companies (corporations) while the second to represent the names of close friends of a person. In such cases, the mappings suggested by a schema matching tool may be incorrect or incomplete. The tool however, may provide some hints to warn the user about potentially incorrect matchings.

The problem of matching specifically ontologies has been studied in many projects, e.g., Knowledge Web and Open Knowledge and a relatively recent survey of the state of the art can be found in [80].

Unfortunately, matching is not enough for semantic inter-operation between two schemas/ontologies. In particular, detailed mappings need to be established between the ontologies/schemas that are to inter-operate. Such mappings can be specified in Datalog or a comparably expressive query language. Clio has been a major project on schema mapping discovery [81] [82], leading to a research prototype developed at the IBM Almaden Research center, as well as a patent [83]. MAPONTO is a related project that focused on the discovery of mappings between schemas and ontologies [84].

Noy [85] has developed a series of tools for performing ontology mapping, alignment and versioning. The last one is the **PROMPT** v3.0, which is a plug-in for Protégé, allowing users to manage multiple ontologies, and in particular to: (i) compare versions of the same ontology, (ii) map one ontology to another, (iii) move frames between included and including project, (iv) merge two ontologies into one, and (v) extract a part of ontology. The tool uses linguistic similarity matches between concepts for initiating the mapping, merging and alignment process, and then uses the underlying ontological structures of the Protégé environment (classes, slots, facets) to inform a set of heuristics for

identifying further matches between the ontologies. Prompt saves the mappings in a mapping ontology developed over the years at SMI (Stanford Medical Informatics) and can perform following complex mapping algorithms: conditional mapping, map of relationships, access values from an instance, perform functional mappings, renaming, splitting and merging of classes, conversions of slots into classes and vice versa.

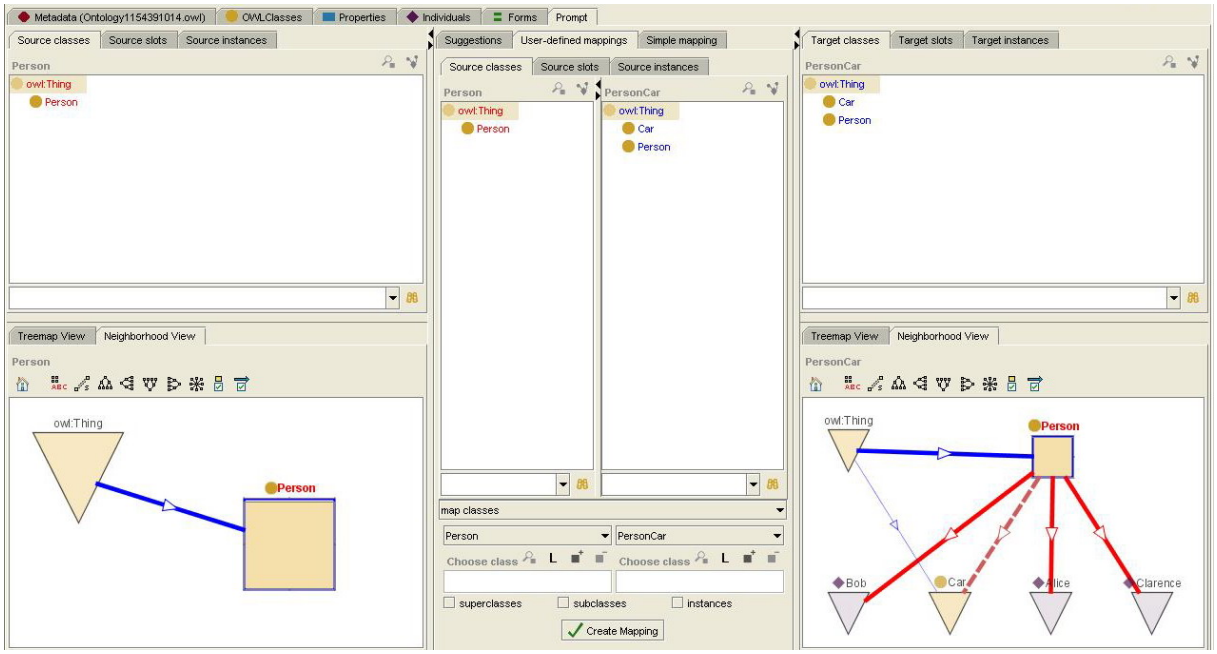

Figure 3-4. The PROMPT Protégé plug-in for mapping between ontologies

3.7. Ontology Visualization

Visualizing the ontologies, either the News or the History ones, in the context of Papyrus is important both for ontology editing and for presentation of the historical information to the end user.

Visualization of ontologies is not a trivial task. An ontology is something more than a hierarchy of concepts. It is enriched with role relations between concepts and each concept has various attributes related to it. Furthermore, each concept most probably has instances attached to it, which could range from one or two to thousands. Therefore, it is difficult to create a visualization that will display effectively all this information and will at the same time allow the user to perform easily various operations on the ontology.

In the field of ontology visualization, there are several works, mostly in 2D. Apart from these systems that propose visualizations especially tailored for ontologies, there are a number of other techniques, used in other contexts such as graph or file system visualization that could also be adapted to display ontologies.

In this section, we present the most prominent techniques among these, with emphasis to the ones that are currently being used in applications. The categorization of the techniques is based on the one provided in [1].

The methods can be grouped according to different characteristics of the presentation, interaction technique, functionality supported or visualization dimensions. For the needs of this survey the methods were grouped in the following categories, representing their visualization type:

1. Indented List
2. Node – link and Tree
3. Zoomable
4. Space – filling
5. Focus + Context or Distortion

Methods grouped in one of these categories may have elements of the other categories, for example, some space-filling techniques may also be zoomable. In these cases the predominant functionality features have been used for the categorization of the method. The effect of possible additional features to the performance of the visualization is presented in the respective discussion section.

This grouping was chosen as a starting point because each of these general categories of visualizations has characteristics that lead to different advantages and weak points. There is a need to investigate how those relate to the special requirements of an ontology visualization tool in relation with the tasks a user would like to perform with an ontology visualization tool.

3.7.1. Indented List

Most of the existing ontology editors, like Protégé[63] [10], OntoEdit [17], Kaon[64] and OntoRama [7], along with their main visualization technique, offer a windows explorer-like tree view of the ontology. In this view, the taxonomy of the ontology (as dictated by the isa inheritance relationships) is represented as a tree (Figure 3-5).

[63] http://protege.stanford.edu

[64] http://kaon.semanticweb.org

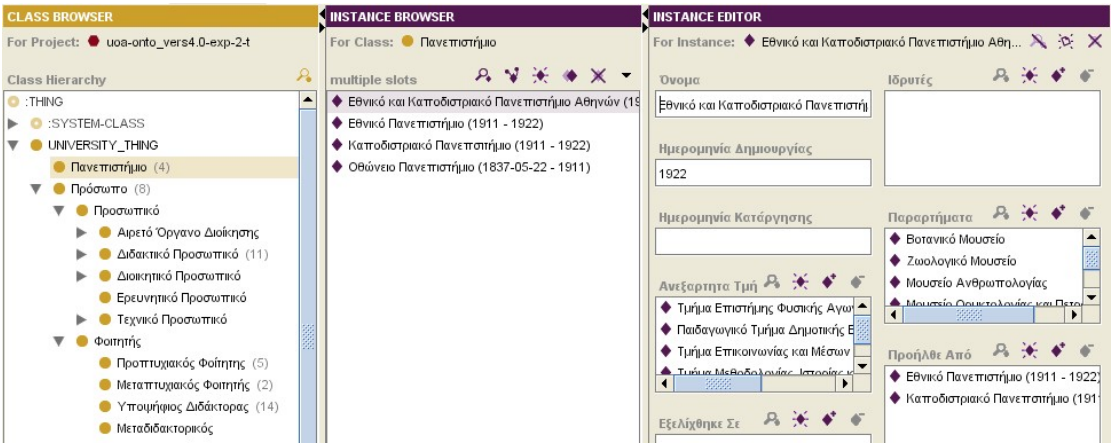

Figure 3-5. Protégé Class Browser

3.7.2. Node – Link and Tree

This category of techniques represents ontologies as a set of interconnected nodes, presenting the taxonomy with a top – down or left to right layout. The user is generally allowed to expand and retract nodes and their sub-trees, in order to adjust the detail of the information shown and avoid display clutter.

An interesting web-based tool that displays related concepts in a node-link style diagram is the SiloBreaker Relationship Network. The user may dynamically adjust the number of concept instances to appear on screen.

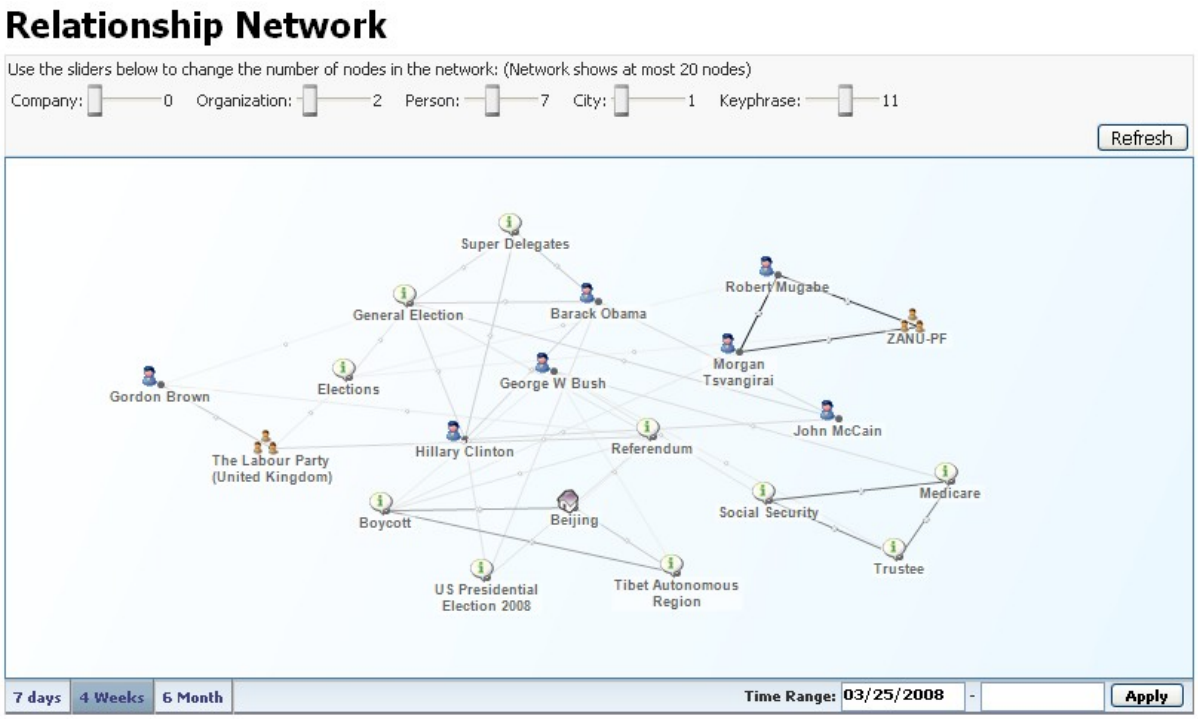

Figure 3-6. The SiloBreaker Relationship Network

OntoViz [13] is a Protégé visualization plug-in using the GraphViz[65] library to create a very simple 2D graph visualization method. The ontology is presented as a 2D graph (Figure 3-7) with the capability

[65] http://www.graphviz.org

for each class to present, apart from the name, its properties and inheritance and role relations. The instances are displayed in different colour. It is possible for the user to choose which ontology features will be displayed, as well as prune parts of the ontology from the Config Panel on the left. Right-clicking on the graph allows the user to zoom – in or zoom – out.

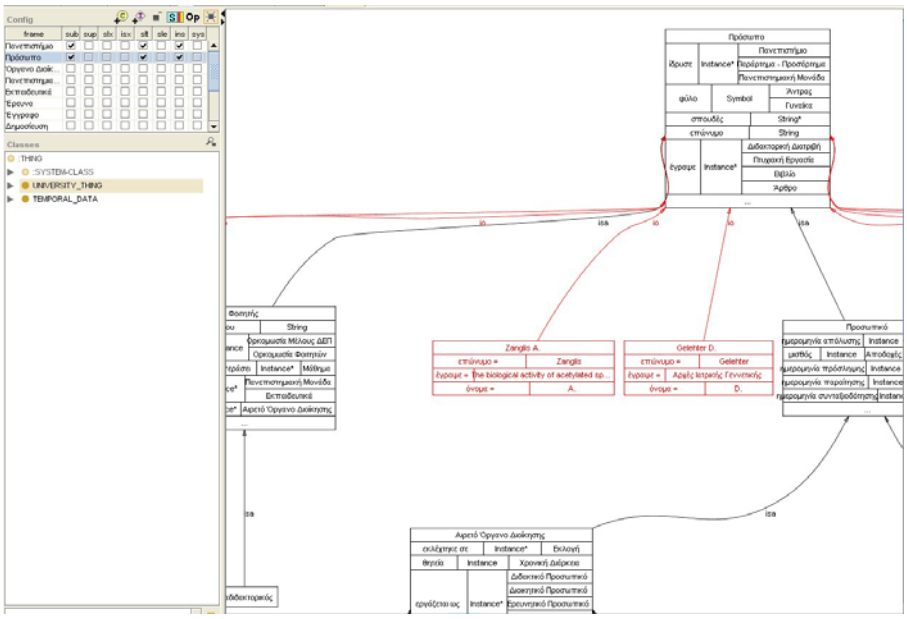

Figure 3-7. Protégé OntoViz visualization

IsaViz[66] is a visual environment for browsing and authoring RDF ontologies represented as directed graphs. Graphs are visualized using ellipses, boxes and arcs between them (Figure 3-8). The nodes are class and instance nodes and property values (ellipses and rectangles respectively), with properties represented as the edges linking these nodes.

OntoTrack[67] is a browsing and editing "in-one-view" authoring tool with a hierarchical layout. It resembles the SpaceTree visualization as it represents retracted sub-hierarchies with triangles of length width and shading that approximates depth, branches and number of sub-classes. As an extra feature, it provides an interface with an external OWL reasoner.

GoSurfer[68] [20], [21] is a data mining tool for visualizing the Gene Ontology[69] (GO) associated with specific genes given as input. It uses a common, top down tree visualization and tools for comparing genes in relation to their corresponding terms in the GO ontology, i.e. comparing ontology paths.

[66] http://www.w3.org/2001/11/IsaViz

[67] http://www.informatik.uni-ulm.de/ki/ontotrack

[68] http://www.gosurfer.org

[69] http://www.go.org

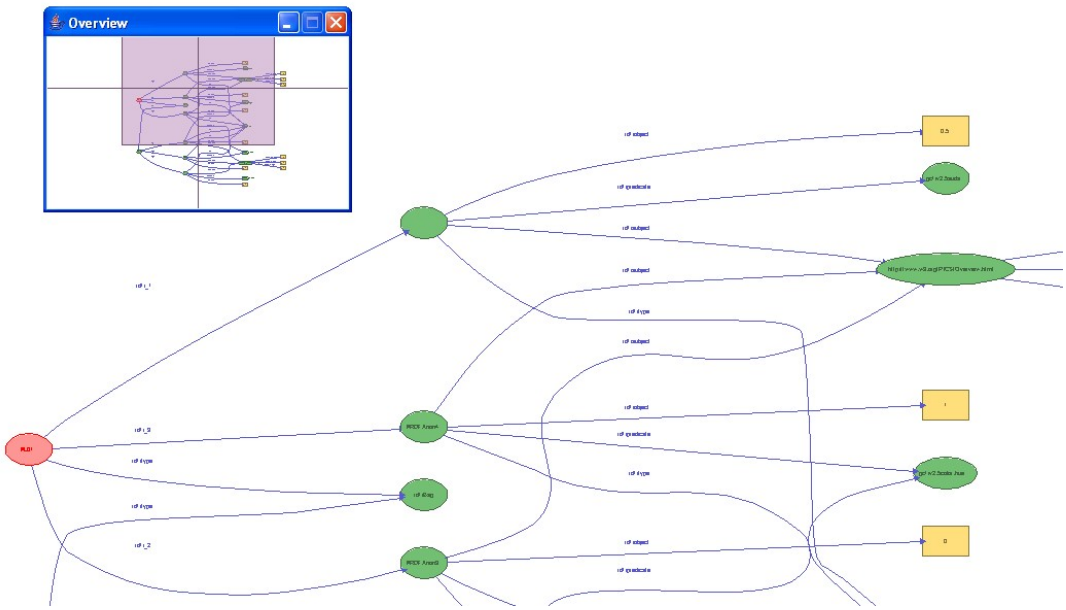

Figure 3-8 IsAviz: Graph with the radar view visible

The **GOBar**[70] visualization [9] is based on the GraphViz library to create an ontology for visualizing GO. **GOMiner**[71] uses a similar top down graph to represent the GO ontology hierarchy.

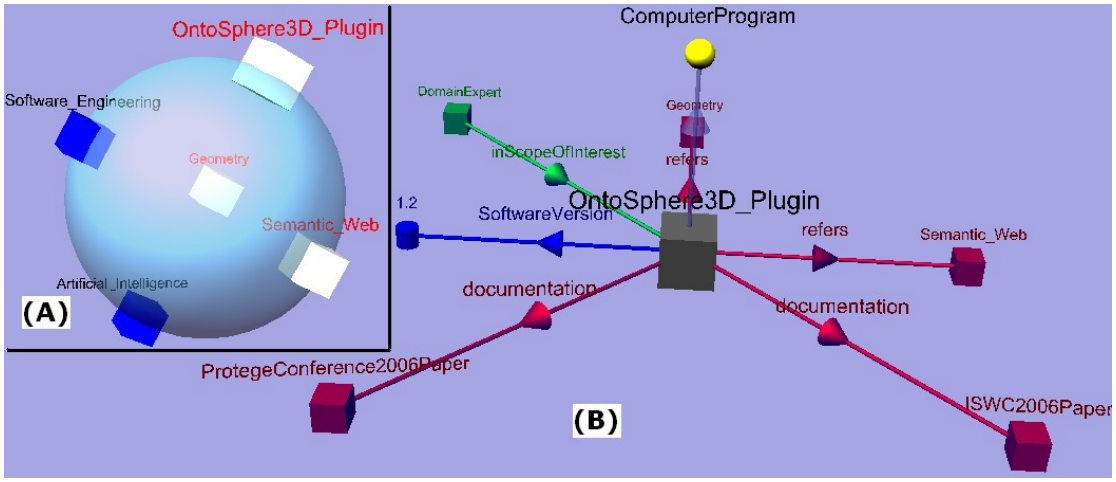

Figure 3-9. OntoSphere visualization (a) Root Focus view (b) TreeFocus view

OntoSphere [5] proposes a node – link tree type visualization that uses three different ontology views in order to provide overview and details according to the user needs. The RootFocus Scene (Figure 3-9.A) presents a sphere bearing on its surface a collection of the upper level classes represented as small spheres. It does not visualize the taxonomy, but the direct role relations between classes. Color and size coding is used to denote existence of sub-trees and their size. The user may right – click on a class to display the RootFocus View of its children. The TreeFocus Scene (Figure 3-9.B), displayed when left–clicking on a class, shows the selected class with its sub-tree. Only three levels down from the selected node are shown expanded. ConceptFocus Scene depicts all the information about the selected class, like ancestors, children and semantic relations.

[70] http://katahdin.cshl.org:9331/GO

[71] http://discover.nci.nih.gov/gominer/

timeViz [22] [23] [24] is a Protégé plug-in that attempts to address the visualization problem for ontologies with temporal characteristics. It is built upon a temporal extension of Protégé [25], which incorporates data types for Date and Period, as well as temporal properties and it focuses mainly on the visualization of the evolution of ontology entities, both classes and instances. It uses the node-link and tree paradigm with a vertical tree representation for the is-a and instance-of relations (Figure 3-10) and a horizontal node-link representation for entity evolution (Figure 3-11).

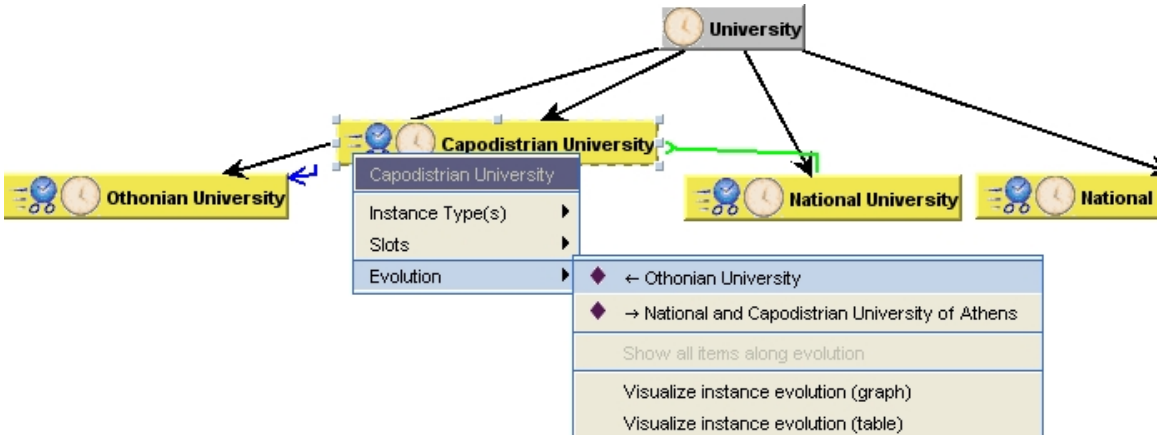

Figure 3-10. Part of the ontology in timeViz, with instance and evolution links visible, as well as the context menu for "Capodistrian University".

The implemented prototype provides users with the following visualization options:

- Restrict the display to classes, entities and relationships pertaining to a specific time period.
- Visualize the entity timeline, i.e. the entity's evolution along the time axis.
- Co-display the timeline of multiple entities

Furthermore, it allows the user to select which ontology sub-hierarchies will be visualized and it provides a high degree of interactivity by the use of context menus. These menus provide a quick overview of the selected entity properties as well as provide options for visualizing related entities or this entity's evolution. (Figure 3-10).

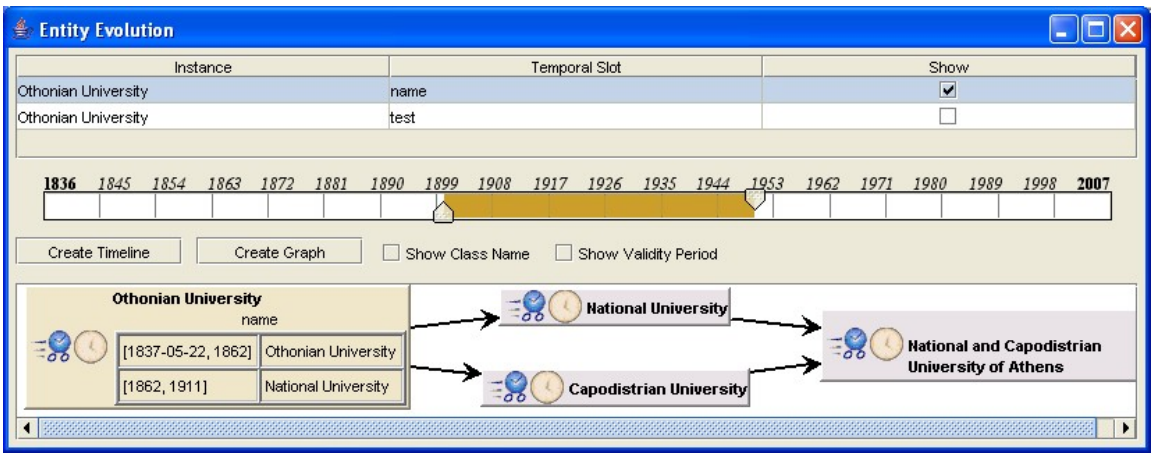

Figure 3-11. Representation of the evolution of the National and Capodistrian University of Athens

3.7.3. Zoomable visualizations

This category contains all the methods that present the nodes in the lower levels of the hierarchy nested inside their parents and with smaller size than that of their parents. These techniques allow the user to zoom-in to the child nodes in order to enlarge them, making them the current viewing level.

Jambalaya [15] is a visualization plug-in for the Protégé ontology tool that uses the SHriMP (Simple Hierarchical Multi-Perspective) [19] 2D visualization technique. SHriMP uses a nested graph view (Figure 3-12) and the concept of nested interchangeable views. It provides a set of tools including several node presentation styles, configuration of display properties and different overview styles.

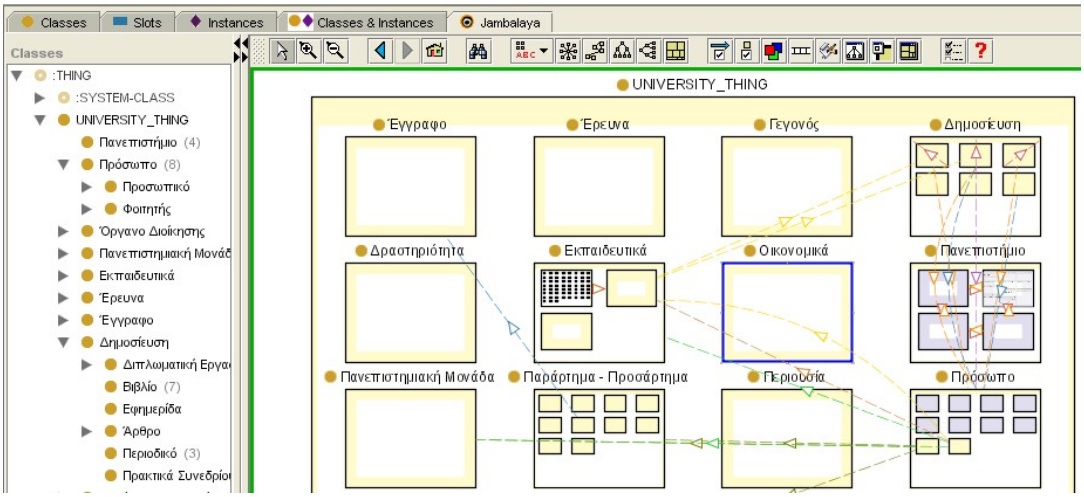

Figure 3-12. The Jambalaya tab in Protégé with Class Browser on the left.

CropCircles[72] [11] [18] is an ontology visualization which represents the class hierarchy tree as a set of concentric circles (Figure 3-13). Nodes are given the appropriate space in order to guarantee enclosure of all the subtrees. If there is only one child, it is placed as a concentric circle to its parents, otherwise the child - circles are placed inside the parent node from the largest to the smallest. The user may click on a circle to highlight it and see a list of its immediate children on a selection pane. The selection pane can let the user drill down the class hierarchy level-by-level and it also supports user browsing history. The user may also select which top level nodes to show in the visualization.

[72] http://www.mindswap.org/2005/cropcircles

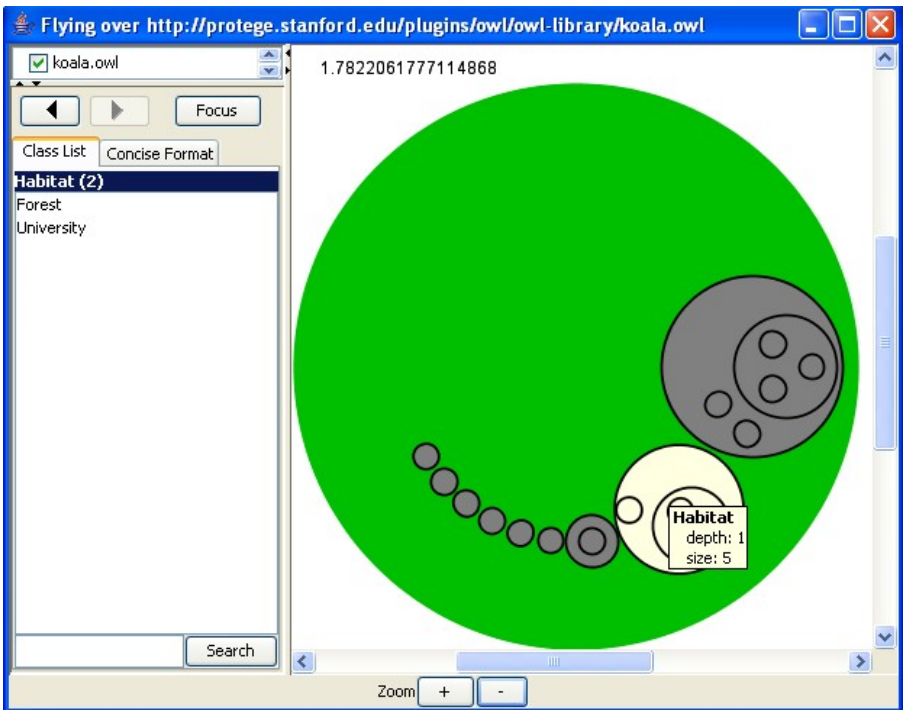

Figure 3-13. TheCropCircles visualization in Swoop. The "Habitat" node is selected and its label visible on mouse over.

3.7.4. Space Filling

Space filling techniques are based on the concept of using the whole of the screen space by subdividing the space available for a node among its children. The size of each sub-division corresponds to a property of the node assigned to it, i.e. its size, number of contained nodes, etc.

The **TreeMaps** [12] visualization method uses a 2D approach of space filling to represent hierarchies, using a rectangular area with rectangular subdivisions (Figure 3-14).

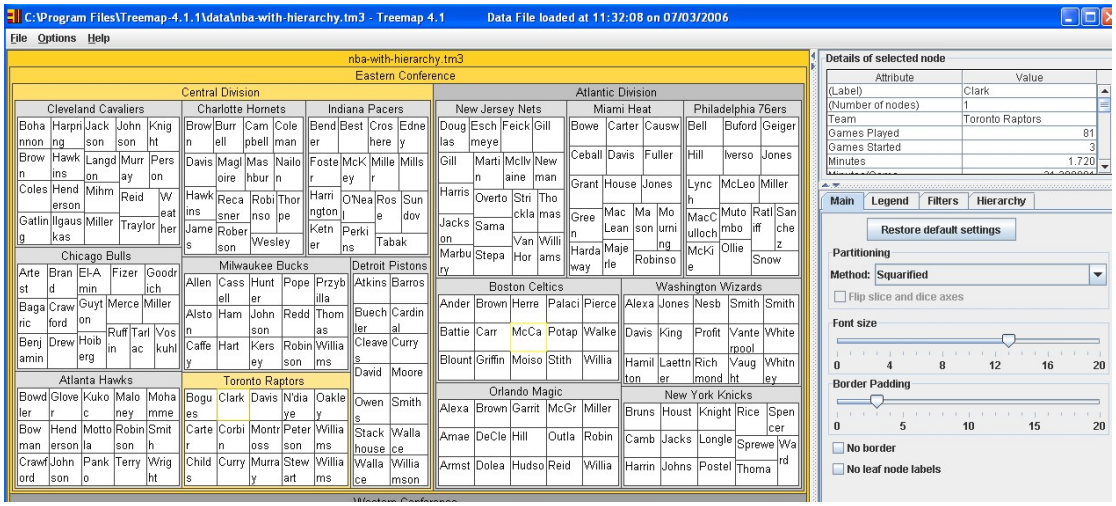

Figure 3-14. Treemap with path to Instance "Toronto Raptors" highlighted

The Treemap technique has been proposed by [4] and [3] as a tool for visualizing the GO ontology. Size and color are used to provide a mechanism to evaluate data. Treemap 4.0 has the functionality to assign labels, size and color to different gene attributes. Moreover, the user may zoom on details by double-clicking on an area of interest so that the area selected is rapidly updated and may query data in the context of the entire GO classification.

3.7.5. Context + Focus and Distortion Techniques

This group of techniques is based on the notion of distorting the view of the presented graph in order to combine context and focus. The node on focus is usually the central one and the rest of the nodes are presented around it, reduced in size until they reach a point that they are no longer visible. Usually a hyperbolic equation is used to this end. The user has to focus on a specific node, in order to enlarge it.

In [14], a 2D hyperbolic tree is used in order to present the ontology of the Brazilian Agricultural Research Society. HyperGraph[73] is an open source project which provides java code to work with hyperbolic geometry and especially with hyperbolic trees. It provides a very extensible api to visualize hyperbolic geometry, to handle graphs and to layout hyperbolic trees.

The hyperbolic tree technique is based on a hyperbolic transformation. The root of the tree is initially placed in the middle of a circular area with the child nodes around it, their child nodes placed around them and so forth. Moving from the centre of the tree to the circumference the distance between the tree levels is diminished in a way that, as a result of the hyperbolic transformation, the whole tree fits

[73] http://hypergraph.sourceforge.net/index.html

in the circular area. The outer nodes, when smaller than a pixel, are not displayed. The technique is therefore based on distortion to keep the visualization within certain limits and combine detailed presentation within the information context. Another commercially available hypertree visualization is the StarTree 74, [8].

OntoRama[75] [5][7] is a Java application used for browsing the structure of an ontology with a hyperbolic – type visualization. Ontorama currently does not support "forest structures", which are sub-hierarchies neither directly nor indirectly connected to the root. It uses cloning of nodes that are related to more than one node, in order to avoid cases where the links become cluttered. It can support different relation types. Apart from the hyperbolic view, it also offers a windows explorer – like tree view.

TGVizTab (TouchGraph Visualization Tab) [2] incorporates the TouchGraph[76] visualization technique in the Protégé ontology management tool. TouchGraph is an open source Java environment for the creation and navigation of network graphs, also employed by the Kaon ontology management tool. It uses a spring – layout technique where nodes repel one another, whereas the edges (links) attract them. This results in placing the semantically similar nodes close to one another. A characteristic of this technique is that it is especially interactive, as the nodes move and adjust to the user commands.

This visualization allows the user to navigate making visible gradually parts of the graph. A variable Radius of visibility is used to limit the size of the graph in smaller, more manageable sizes. The user may also expand or retract nodes, hide them and change the node on focus by double clicking on it. Furthermore, s/he has full control on the color and visibility of the links and may change the zoom level or make the graph hyperbolic.

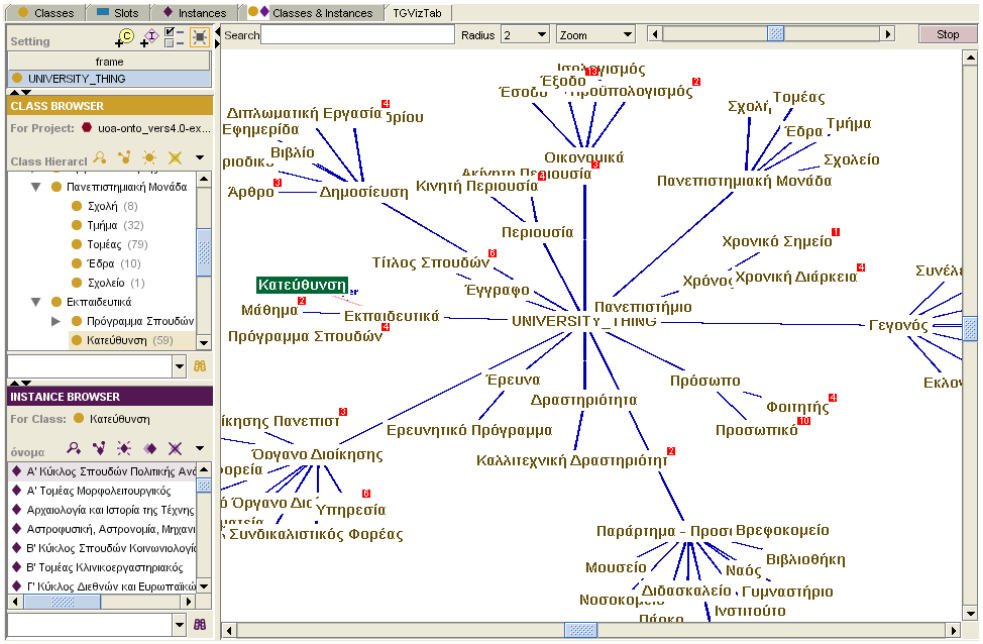

Figure 3-15. Protégé TGVizTab

[74] http://www.inxight.com/

[75] http://www.ontorama.com

[76] http://www.touchgraph.com/

Figure 3-15 presents the interface of the TGVizTab. The ontology is presented as a tree structure on the left (Class Browser). In order to create the visualization on the right, a class or instance should be selected as a starting focal point.

OZONE (Zoomable Ontology Navigator) [16] is a visual interface for searching and browsing ontological information. OZONE visualizes query conditions and provides interactive, guided browsing for DAML (DARPA Agent Markup Language) ontologies. OZONE reads ontology information and rearranges it visually with context information so that ontology information can be queried and browsed easily without knowledge of their structure. Queries can be formulated interactively and incrementally by manipulating objects on the screen.

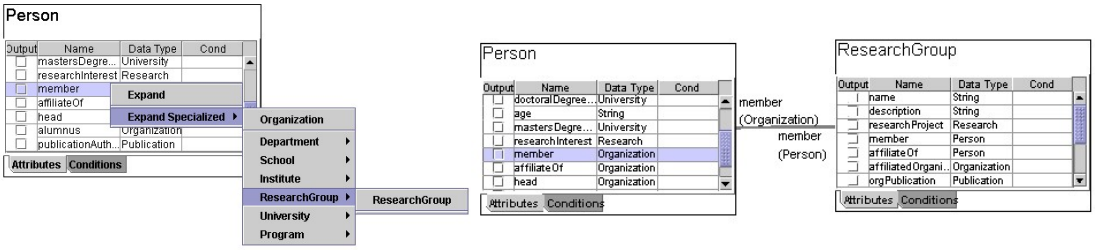

Figure 3-16. Selecting a property (left) and the expanded node (right)

For example, if the user wants information about people, s/he begins to form a query by selecting the "Person" class from a class list that contains all classes of the ontology. This action puts the "Person" class on the display. Since the goal of the query is to find information about people in a particular research group, the user scans the properties of the "Person" to find a property that relates a person with an organization (Figure 3-16). The user clicks the "member" property of the visual node because s/he finds that it is the most appropriate property to specify "is a member of" relationship. When the user clicks, a pop-up menu appears.

In OZONE, any sub-graph can be grouped and transformed into a single node by choosing the 'Group' menu in the main menu after selecting nodes on the screen. The collection of nodes is zoomed out and a simple new node replaces the collection. The user can access the detailed sub-nodes at any time by zooming in.

3.7.6. Conclusions

In previous sections we presented the main ontology visualization types as they are proposed in existing ontology editing tools and research. The following table summarizes these results.

Table 3-1. Existing ontology visualizations

Visualization Type	Method	Software availability
Indented List	Protégé Class Browser[77]	Open Source
Node Link and Tree	OntoViz[12]	Open Source, available as a Protégé plug-in
	GOBar[78]	Freely available as a web – based tool
	IsAviz[79]	Open source, available, Possibility to create plug-ins
	OntoTrack[80]	Available under non commercial license

[77] http://protege.stanford.edu

[78] http://katahdin.cshl.org:9331/GO

[79] http://www.w3.org/2001/11/IsaViz

[80] http://www.informatik.uni-ulm.de/ki/ontotrack

Visualization Type	Method	Software availability
	GOSurfer[81]	Freely available
	Ozone	No
	timeViz[82]	Open Source, available as a Protégé plug-in
	GOMiner[83]	Freely available
	OntoSphere[84]	Available as a Protégé plug-in in
Zoomable	Jambalaya[12]	Open Source, available as a Protégé plug-in
	CropCircles[85]	Preliminary Java version available
Space - filling	TreeMap 4.0[86]	Available commercially and as a free demo version
Context+focus and Distortion	HyperTree Visualization	No
	OntoRama[87]	No
	StarTree[88]	Available as part of a commercial application
	TGVizTab[12]	Open Source, available as a Protégé plug-in

The visualization of ontologies is a particular sub-problem of the field of graph and hierarchy visualization with many implications due to the various features that an ontology visualization should present. As [1] implies, there is not one specific method that seems to be the most appropriate for all applications and, consequently, a viable solution would be to provide the user with several visualizations, so as to be able to choose the one that is the most appropriate for his/her current needs. Some ontology management tools already provide combinations of visualization methods. Protégé for example includes several visualization plug-ins that are coupled with the Protégé indented list Class Browser.

Furthermore, an important conclusion of most of the evaluations taken into account in [1] is that visualizations should be coupled with effective search tools or querying mechanisms. Browsing is not enough for tasks related to locating a specific class or instance, especially for big ontologies. Most users also seem to dislike chaotic and over-cluttered overviews and tend to prefer visualizations that offer the possibility of an orderly and clear browsing of the presented information, even if in some cases it requires focusing on a specific part of the ontology or hierarchy. This fact implies that visualizations should also take advantage of the semantic context of the information and even the user profile in order to guide and support the hierarchy or ontology exploration.

In some applications it is preferable or more convenient to provide only a single visualization of the ontology. In this case the designer has to make a choice between the available methods, based on certain characteristics of the ontology, the application, the user profile and expertise and so forth.

[81] http://www.gosurfer.org

[82] http://oceanis.mm.di.uoa.gr/pened/index.php?c=publications#plugins

[83] http://discover.nci.nih.gov/gominer/

[84] http://ontosphere3d.sourceforge.net/

[85] http://www.mindswap.org/2005/cropcircles

[86] http://www.cs.umd.edu/hcil/treemap

[87] http://www.ontorama.com

[88] http://www.inxight.com/

In the case of Papyrus, user needs and requirements will be taken into account in order to select the appropriate approach to be used for ontology visualization, either for editing or for browsing for the end user. The representation of time in the context of ontologies is an issue of special importance in the context of this project, as already mentioned.

4. Multimedia Content Analysis

The objective of research in the area of Content Analysis in the context of Papyrus is to develop tools for knowledge based interpretation of multimedia content from news archives in the context of historical knowlege mapping. Recent research in multimedia content analysis focus on narrowing the gap between low-level content descriptions that can be computed automatically by a machine and the richness and subjectivity of semantics in high-level human interpretations of audiovisual media, commonly known as "Semantic Gap". The notion of semantic gap can be interpreted in different ways depending on the content. Thus in Papyrus reseach will focus on developing techniques to derive high level interpretation of multimedia content from news archives. To enhance multimedia content analysis from news arhcives, a multimodal approach will be followed integrating audio, image, video and text analysis. Multidisciplinary research aspects of content analysis will be investigated including content structuring, pattern recognition, event detection, audio processing, text analysis and data mining. The tools will investigate different multimedia content processing techniques based on evolutionary algorithms, fuzzy systesms, biologically inspired techniques and kernel-based machine techniques for low-level content analysis. One of the key research aspect focuses on knowledge assisted multimedia analysis for extraction of semantically meaning objects, events and content structure from news archives. To achieve an automatic transition from low-level features to symbolic entities, different segmentation alogirthms and pattern recognition techniques will be employed based on MPEG – 7 low-level features. In the context of Papyrus, intelligent machine learning techniques will be developed investigating neurofuzzy networks, support vector machines and biologically inspired techniques.

In this chapter, an overview on state of the art multimedia content analysis techniques within the context of Papyrus is presented. The reminder of the section is organised as follows. In section 4.1 an overview of the MPEG – 7 visual features is presented with specific focus on colour and texture descriptors. In section 4.2 multimedia content structuring tools are presented with emphasis on video (visual) and audio based algorithms, followed by techniques for intelligent relevance feedback in section 4.3. Section 4.4 presents the techniques for knowledge assisted analysis techniques for extracting high-level semantic concepts from low-level features. In section 4.5, an overview of the mutimodal techniques for integrating audio-visual features are presented followed by text retrieval in Section 4.6.

4.1. MPEG – 7 Visual Features

Recent years have seen a rapid increase in volume of image and video collections. A huge amount of information is available and every day gigabytes of new visual information is being generated, stored and transmitted. However, it is difficult to access this visual information unless it is organised in a suitable way – to allow efficient browsing, searching and retrieval. A very popular means for image and/or video retrieval is to annotate images or video with text, and to use text-based database management systems to perform image/video retrieval. However, text-based annotation has significant drawbacks when confronted with large volumes of images/videos. Manual annotation can in these circumstances become significantly labor intensive. Furthermore, since images are rich incontent, text in many applications maynot be rich enough to describe images. Hence, to overcome these difficulties in early 1990s, content based image retrieval emerged as a promising means for describing and retrieving images. Content – bsed image retrieval systems describe images by their own visual content rather than text, such as colour, texture and object shape information. For this prupose, in 1997 the ISO MPEG Group initiated the "MPEG – 7 Multimedia Description Language" work item. The goal and objective of MPEG – 7 Visual Standard is to provide standardised descriptions of streamed or stored images or video – standardised header bits that help users or applications to identify, categorise or filter images or video. These low-level descriptors can be used to compare, filter or browse images or video purely based on nontext visual descriptions of the content. In the following section, an overview of the MPEG – 7 visual features are presented. The presented visual descriptors will be further revisited in Sections 4.3 and 4.4 while Presenting techniques for intelligent relevance feedback and knowledge extraction from multimedia content.

Colour Descriptor

Colour and texture are the dominant parameters for human perception. In this section the main focus will be on colour descriptor. Due to the compact representation and low complexity, colour descriptors are preferred in applications like image retrieval. However, it has many serious drawbacks those including a high degree of dependency on colour codebook design, sensitivity to quantization boundaries, and inefficient in representing images with few dominant colours [100]. The MPEG – 7 colour descriptors consist of a number of histogram descriptors, a dominant colour descriptor (DCD) and a colour layout descriptor (CLD) [101].

Colour Structure Descriptor

The colour structure descriptor represents the colour distribution of a content (as in colour histogram) and local spatial structure of this content. The colour structure descriptor is constructed by scanning a colour quantized image with 8 X 8 structure window and generating a colour histogram (HMMD colour space) with predefined number of bins.

The CSD is identical in form to a colour histogram but is semantically different. Specifically, the CSD is an $1D$ array of eight bit – quantized values.

$$CSD = \overline{h}_s(m), m \in \{1,2,...,M\} \tag{4.1.1}$$

Where M is chosen from the set $\{256,128,64,32\}$ and where s is the scale of the associated square structuring element. The CSD uses the $L1$ norm for matching in its similarity measure based on the retrieval results the HMMD colour space is non-linearly quantized. The CSD provides information regarding colour distribution as well as localized spatial colour structure in an image. The image is represented by a modified colour histogram that incorporates the spatial distribution of each colour into the description. The distance between two CSD histograms for images Q and I is calculates using the $L1$ norm as follows.

$$D_{CSD}(Q,I) = \sum_{i=0}^{127} | H_{Q,i} - H_{I,i}) | \tag{4.1.2}$$

Where $H_{Q,i}$ represents the i^{th} bin of the colour structure histogram for image Q and a 128 bin histogram has been described in [106].

Dominant Colour Descriptor

The descriptor specifies local colour features, enabling compact and effective description of representative colours in a region or an image. These colours are computed and quantized for each region/image and are not fixed in the colour space [107] [108].

$$F = \{(c_i, p_i, v_i), s_i\},\ i = 1,2,3,...,N \tag{4.1.3}$$

Where N is the number of dominant colours; c_i is dominant colour value vector with corresponding colour components values; p_i is percentage of pixels that have corresponding colour values in the region; v_i is the variance describing variation of colour values for pixels in a cluster of a particular representative colour; s_i is the spatial coherency defining the homogeneity of dominant colours in the image. Typically three to four colours provide a good characterization of the region colours. In similarity matching, searching for individual colours can be done very efficiently in a 3 – D colour space. The dissimilarity measure for two descriptors' is given below.

$$D^2(F_1,F_2) = \sum_{i=1}^{N1} p^2{}_{1i} + \sum_{j=1}^{N2} p^2{}_{2j} - \sum_{i=1}^{N1}\sum_{j=1}^{N2} 2a_{1i,2j}\ p_{1i}p_{2j} \tag{4.1.4}$$

Where the subscripts 1 and 2 in all variables stand for descriptions F_1 and F_2, respectively and a_{kl} is the similarity coefficient between two colours c_k and c_l.

$$a_{k,l} = \begin{cases} 1 - d_{k,l} / d_{\max}, d_{k,l} \le T_d \\ 0, d_{k,l} > T_d \end{cases} \quad (4.1.5)$$

Where $d_{k,l} = \| c_k - c_l \|$ is the Euclidean distance between two colours c_k and c_l, t_d is the maximum distance for two colours to be considered similar and $d_{\max} = a.T_d$. In particular, this means that any two dominant colours from one single description are at least T_d distance apart. One variation of the above distance is to use the spatial coherence field.

$$D_s = \omega_1 abs(s_1 - s_2)D + \omega_2 D \quad (4.1.6)$$

Where s_1 and s_2 are the spatial coherencies of the query and target descriptors and ω_1 and ω_2 are the fixed weights, with recommended settings to 0.3 and 0.7 respectively. If the colour variance is considered, the matching function is based on modelling of the colour distribution as a mixture of Gaussian distributions with parameters defined as colour values and colour variances. Calculation of the squared difference between the query and target distributions then leads to the following formula for the matching function.

$$D_v = \sum_{i=1}^{N1}\sum_{j=1}^{N2} p_{1i}, p_{1j} f_{1i1j} + \sum_{i=1}^{N1}\sum_{j=1}^{N2} p_{2i}\, p_{2j}\, f_{2i2j} - \sum_{i=1}^{N1}\sum_{j=1}^{N2} 2\, p_{1i}\, p_{2j}\, f_{1i2j} \quad (4.1.7)$$

$$fxiyj = \frac{1}{2\pi\sqrt{v_{xiyj}^{(l)} v_{xiyj}^{(u)} v_{xiyj}^{(v)}}} *$$
$$\exp[-\left(\frac{c_{xiyj}^{(l)}}{v_{xiyj}^{(l)}} + \frac{c_{xiyj}^{(u)}}{v_{xiyj}^{(u)}} + \frac{c_{xiyj}^{(v)}}{c_{xiyj}^{(v)}}\right)/2] \quad (4.1.8)$$
$$c_{xiyj}^{(l)} = (c_{xi}^{(l)} - c_{yj}^{(l)})^2$$
$$v_{xiyj}^{(l)} = (v_{xi}^{(l)} - v_{yj}^{(l)})$$

In the equation above, $c_{xi}^{(l)}$ and $v_{xi}^{(l)}$ are dominant colour values and colour variances, x and y index the query and target descriptors, i, j index the descriptor components and l, u and v the components of the colour space.

Colour Layout Descriptor

The Colour Layout Descriptor (CLD) is a compact and resolution invariant representation of colour for high speed image retrieval. The descriptor is designed to capture the spatial distribution of colour in an image or an arbitrary shaped region. The spatial distribution of colour constitutes an effective descriptor for sketch based region image retrieval, content filtering using image indexing, and visualization. The functionality of this descriptor can also be achieved using a combination of grid structure descriptor and grid-wise dominant colours. However, such a combination would require a relatively large number of bits, and matching will be more complex and expensive. The CLD uses representative colours on a $8X8$ grid followed by a DCT and encoding of the resulting coefficients. The feature extraction process consists of two parts; grid based representative colour selection and the DCT transform with quantization. The DC values are quantized to 6 bits and the remaining to 5 bits each. These results demonstrate that the CLD is quite effective in image retrieval. The results also compare favourably with a grid based dominant colour approach wherein the image is partitioned and dominant colours for these partitions are used to represent the layout. For matching two CLD's $\{DY, DCr, DCb\}$ and $\{DY', DCr', DCb'\}$, the following distance measure can be used:

$$D = \sqrt{\sum_i \omega_{yi}(DY_i - DY'_i)^2} + \sqrt{\sum_i \omega_{bi}(DCb_i - DCb'_i)^2} + \sqrt{\sum_i \omega_{ri}(DCr_i - DCr'_i)^2} \quad (4.1.9)$$

In the distance equation i represents the zigzag scanning order of the coefficients. The perceptual characteristic of human vision system could be included for similarity calculation since the feature description is in frequency domain. The distance is weighted appropriately, with larger weights given to the lower frequency components, to match the characteristics.

An important problem in colour-based image retrieval and video segmentation is to lack information regarding spatial distribution of colour. To solve this problem and enhance the performance of image and video analysis, a spatial colour descriptor is proposed involving adjacency histogram and colour vector angle histogram in [102]. A similarity metric is needed when using the proposed spatial colour descriptor for image retrieval and video cut detection. When measuring the similarity of a colour adjacency histogram, the intersection method is used, which measures the similarity of the binary codes for the same colour between the query and model images. Let $B_{Ck}(I) = (b_1^k, b_2^k, ..., b_n^k)$ denote the binary code of adjacent colours to colour Ck in query image I, then the intersection result of query image I and model image J concerning colour Ck is calculated as

$$S_k(I,J) = \frac{NC_k^1(I,J)}{N_k^1(I) + N_k^1(J) - N_k^1C(I,J)} \tag{4.1.10}$$

Where $N_{k,l}$ the total number of binary 1 in a binary string and $NC_{k,l}$ is the total number of binary 1's occurring at the same position in the two binary strings. For all n colours used in the construction of a colour adjacency histogram, the total intersection is computed as

$$S_{pair}(I,J) = \frac{1}{n}\sum_{k=1}^{n} S_k(I,J) \tag{4.1.11}$$

When measuring the similarity of a colour vector angle histogram, Swain's histogram intersection is used. The total number of pixels in a colour vector angle histogram of an image is dependent on the image size and number of edge pixels in each image. The similarity of a colour vector angle histogram with the same number of pixels is calculated as

$$S_{vec}(I,J) = 1 - \sum_{k=1}^{n} | h_{veci}(k) - h_{vecj}(k) | \tag{4.1.12}$$

Where $j_{vec}(k)$ and $h_{vec}(k)$ represent the colour vector angle histograms of the query and model images, respectively. Based on the similarity measure, the overall similarity measure is represented as

$$S(I,J) = \alpha * S_{pair}(I,J) + \beta * S_{vec}(I,J) \tag{4.1.13}$$

Where α and β are the similarity weights of the colour adjacency histogram and colour vector angle histogram, respectively. Given a video stream, V composed of N frames $\{f_i\}$, the sequence trace can be defined as follows. Let $x_i = \{x_{1i}, x_{2i}, ..., x_{ni}\}$ be the feature set extracted from the pair of frames $\{f_i, f_{i+1}\}$. The sequence trace, d_i for V is defined as

$$d_i = \frac{1}{2T}\sum_{k=1}^{n} x_{ki} \tag{4.1.14}$$

Where T is the total number of pixels in each frame, also the colour adjacency histogram and colour vector angle histogram are used as the features; hence $n = 2$. The first feature x_{1i} is obtained from the absolute difference in the colour adjacency histograms between frames f_i and f_{i+1}

$$x_{1i} = \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} | h_{adj_i}(k,l) - h_{adj_{i+1}}(k,l) | \tag{4.1.15}$$

Where h_{vec_i} and $h_{vec_{i+1}}$ are the colour vector angle histograms of frames f_i and f_{i+1} respectively, and N is the number of representative colours used in the construction of the colour vector angle histograms. If the sequence trace d_i is larger then a predetermined threshold, a scene cut is declared.

Texture Descriptors

Image texture has emerged as an important visual primitive for searching and browsing through large collections of similar looking patterns. The texture descriptors in MPEG – 7 facilitate browsing and similarity retrieval in image and video databases. The three texture descriptors of texture are Homogenous Texture Descriptor (HTD), Texture Browsing Descriptor (TBD) and Edge Histogram Descriptor (EHD).

Homogenous Texture Descriptor

The homogenous texture descriptor provides a quantitative representation using 62 numbers, consisting of the mean energy and the energy deviation from a set of frequency channels. At first, the image is filtered with a bank of orientation and scale sensitive filters [103], [104] and [105]. The frequency domain is partitioned into 30 channels, modelled by 2D – Gabor functions. The energy and standard deviation of each channel are computed, and logarithmically scaled to obtain e_i and d_i respectively. Where I is the i^{th} feature channel, mean and standard deviation of the whole image are denoted as f_{DC}, f_{SD} . The descriptor is formed as

$$HTD = [f_{DC}, f_{SD}, e_1, ..., e_{30}, d_1, ..., d_{30}] \tag{4.1.16}$$

The similarity distance between two texture images is measured by summing the weighted absolute difference between two sets of feature vectors, one for a query image (TD_{query}) and the other in the database $(TD_{database})$. Similarity is measured by calculating the distance between two feature vectors, and is given by

$$\begin{aligned} &d(TD_{query}, TD_{database}) = dist(TD_{query}, TD_{database}) \\ &d(TD_{query}, TD_{database}) = \sum_k \left| \frac{TD_{query}(k) - TD_{database}(k)}{\alpha(k)} \right| \end{aligned} \tag{4.1.17}$$

For intensity invariance matching, the first component of the feature vector, the average intensity of the image, is not used in computing the dissimilarity. The texture descriptor of a rotated image is an angular – shifted version of the reference domain. The distance between texture vectors in the database and a query texture vector by shifting the query texture vector in angular direction is measured as follows.

$$d(TDquery, TDdatabase, m\phi) = dist(TDquery, TDdatabase) \tag{4.1.18}$$

Where ϕ = 30 degrees. Then, for rotation invariant matching the distance is calculated as follows.

$$d(TD_{query}, TD_{database}) = \arg\min_m \{d(TD_{query}, TD_{database}, m\phi) \mid m = 0, ..., 5\} \tag{4.1.19}$$

A promising direction for this work combines the discriminating capability of the texture descriptor with the land – type labels from an extensive gazetteer in supervised learning framework.

Texture Browsing Descriptor

The texture browsing descriptor (TBD) specifies the perceptual characterization of a texture, which is similar to a human characterization, in terms of regularity, coarseness and directionality. The texture is represented as follows.

$$TBD = [v_1, v_2, v_3, v_4, v_5] \tag{4.1.20}$$

v_1 represents how regular or structured the texture is. The larger the value of v_1, the more regular is the texture. Two quantized directions that best capture the directionality of the texture. Two quantized scales that best capture the coarseness of the texture. A larger value indicates a coarser

texture. The descriptor is extracted by filtering the image with Gabor filters. From multi resolution decomposition, a given image is decomposed into a set of filtered images. Each of these images represents the image information at a certain scale and at a certain orientation. Its computation is based on the following observations.

- Structured textures usually consist of dominant periodic patterns.
- A periodic or repetitive pattern, if it exists, can be captured by the filtered images. This behaviour is usually captured in more than one filtered output.
- The dominant scale and orientation information can also be captured by analyzing projections of the filtered images.

Edge Histogram Descriptor

Spatial distribution of edges in an image is another useful texture descriptor for similarity search and retrieval. The edge histogram descriptor (EHD) represents local edge distribution in the image. Specifically dividing the image space into $4X4$ sub-images, the local edge distribution for each sub-image can be represented by a histogram. To generate the histogram, edges in the sub-images are categorized into five types; vertical, horizontal, 45' diagonal, 135' diagonal and non-directional edges. Since there are t_{16} sub-images, a total of 16 * 5 = 80 histogram bins are required. For matching purposes, in some applications, the local edge histograms alone may not be sufficient for an effective image matching. Specifically, edge distribution information for the whole image space and some horizontal and vertical semi-global edge distributions as well as local ones are required to improve the matching performance. The global and semi global edge histograms are directly computed from the 80 local histogram bins $h(i)$, $i = 0,1,2...,79$. For similarity matching, the $L1$ distance measure $D(a,b)$ can be adopted for two image histograms A and B as in the following equation.

$$D(A,B) = \sum_{i=0}^{79} | h_A(i) - h_B(i) | + 5 * \sum_{i=0}^{4} | h_A{}^g(i) - {}_b^g h(i) | + \sum_{i=0}^{64} | h^S{}_A(i) - h^S{}_B(i) | \qquad (4.1.21)$$

Where $h_A(i)$ and $h_B(i)$ represents the normalized histogram bin values of images A and images B, respectively. $h_A^g(i)$ and $h_B^g(i)$ represents the normalized bin values for the global-edge histograms of image A and image B respectively, which are obtained from the corresponding local histograms $h_A(i)$ and $h_B(i)$, $h_A^s(i)$ and $h_B^s(i)$ represents the histogram bin values for the semi-global edge histograms of image A and image B, respectively. Since the number of bins of the global histogram is relatively smaller than that of local and semi-global histograms, a weighting factor of 5 is applied.

4.2. Content Structuring tools

Structural analysis of video is a prerequisite step to automatic video content analysis. The challenge of multimedia content structure modelling corresponds to the task of developing mathematical representations of audio-video content structure to semantic multimedia concepts. However, the recent escalation of audio-visual activity serves greatly to illuminate the deficiency in video modelling solutions. In response to this, there currently exists an abundance of research projects worldwide that aim to provide solutions to many of the aspects of the video structure-modelling question. For instance, much attention has recently been paid to the task of making the WWW more searchable for multimedia content, exemplified by the ISO/IEC standard MPEG-7. The MPEG-7 standard aims to tackle the issue by offering a comprehensive set of low-level audiovisual description tools for creating descriptors, facilitating search, filtering and browsing activities. However, experiential evidence suggests that users of content collections prefer to query video content at the conceptual or semantic level rather than at a feature level [109] - hence the issue of the '*semantic gap*' in video processing [110]. The "*semantic gap*" is a multimedia retrieval-based concept that relates to the virtual gap between the rich meaningful intelligence that a user desires (e.g. scene segmentation, genre recognition, event detection, summarization, etc.) and the shallowness of the multi-modal description features that may be automatically extracted from the content (e.g. colour, edge, texture, motion, audio level/pitch, etc.). It is a commonly held principle among the research community, that one of

the most pressing aspects of the video modelling question concerns this issue of extending the nature of user interaction with multimedia content towards real semantics. That is, bridging the semantic gap may be seen as the fundamental challenge to be overcome in the development of most real-world video structure-modelling applications. In the following section, different video modelling techniques such as shot boundary, scene detection are presented using visual and audio features. These techniques will be used to extract the high level semantics and grammar of the multimedia content such as "interviews", "specific event detection", etc.

4.2.1. Visual

Shot Boundary Detection

Shot boundary detection is one of the common techniques for content structuring. This section presents a representative selection of the leading approaches in this area. As it is a fundamental problem, much of the early research on video analysis concentrated on this task. As 'hard' cuts (i.e. when the video moves from one shot to another in successive frames) is the most common shot boundary type, many techniques focus specifically on detecting this type of shot cut. Boreczky et al. [111] identify the main techniques in shot boundary detection as pixel difference, statistical difference, histogram comparison, edge difference, compression difference and motion vector based. Each of these techniques were implemented, and extensively tested on a dataset consisting of general television programs, news, movies and commercials. In general, it was found that the simpler algorithms (specifically colour histogram based algorithms) performed best. Browne et al. [112] also implemented three different shot boundary detection methods: a colour histogram based approach, an edge histogram based approach, and an approach using the motion information encoded in the macroblocks of MPEG-1 data. The three approaches were tested on thirteen different videos, ranging from news to cookery programs. Overall, it was found that a colour histogram approach outperformed the edge detection and macroblock based approaches.

The TRECVID[89] series of workshops evaluates different approaches in video retrieval by providing a common video dataset, as well as a set of tasks, to participants. Participants complete the required tasks, and submit their results. As there is a common video data set, a direct comparison of different techniques can be made. For the past number of years, TRECVID has run a shot boundary detection task [113]. In this task, a common set of video (most recently, news content in 2005) is supplied to participants, and they are required to submit a set of automatically detected shot boundary times. Many of the techniques in this shot boundary task are now evolving to detect long fades of dissolves, as the established methods of detecting hard cuts achieve very accurate results. The approach of Petersohn et al [114] was one of the more successful techniques in TRECVID. This technique focuses on detecting three types of shot transitions, hard cuts, dissolves/fades and wipes. Each of the shot transitions are detected independently using colour, motion and edge data, and the results are combined to produce a final set of shot boundaries. Similarly, Volkmer et al. [115] use a moving window throughout the frames in a video, and detect gradual transitions by comparing the colour information of all frames in the window. The use of a sliding window means that it is possible to map the changes between frames over a long period of time, rather than just from frame to frame. As gradual transitions typically take a large number of frames, the authors claim that this approach is more suitable.

For many applications the selection of a representative frame for each shot (known as a keyframe) is straightforward. Any of the frames contained within the shot could be used as a keyframe. Many applications choose the keyframes temporally, by picking the first, last, or middle frames of a shot. However, there are more sophisticated approaches to keyframe selection. Vermaak et al. [116] aim to select keyframes that are distinct from every other frame in a shot, and therefore, the authors claim that different frames carry different information. Hence, the authors proposed to use colour information to analyse frame dissimilarity for extracting keyframe. Cooper et al [117] extract keyframes with the aim of only extracting keyframes that are dissimilar to other frames, thus ensuring that much information is presented in a relatively low amount of images. DCT coefficients are used to

[89] http://www-nlpir.nist.gov/projects/trecvid/

discriminate between frames. Girgensohn et al [118] extract keyframes from entire videos, rather than individual shots, by examining the temporal length of the shots, as well as clustering information based on colour similarity.

Scene Segmentation

Detecting scene boundaries enhances the known structure of a video, as rather than a video-shot representation, a more representative video-scene-shot structure is available. Yeung and Yeo [119], [120] proposed a technique in which time constrained clustering of shots is used to build a scene transition graph. This involves grouping shots that have a strong visual similarity and are temporally close, and based on these clusters the authors identify the scene transitions. The clustering is threshold-based and any number of clusters can be created. Hence, scene boundaries are located by examining the structure of the clusters and detecting points where one set of clusters ends, and another begins. The approach presented by Rui et al in [121] also aims to group related shots together based on colour and activity histograms, and then detect the scene boundaries based on the shot groupings.

Kender et al [122] utilise the idea of shot coherence in order to find scene boundaries. Instead of clustering similar shots together, the coherence is used as a measure of the similarity of a set of shots with previous shots. The coherence is based on colour similarity using histograms. When there is 'good coherence', many of the current shots are related to the previous shots and therefore judged to be part of the same scene, when there is 'bad coherence', most of the current shots are unrelated to the previous shots and a scene transition is declared. Rasheed et al. [123] again uses the concept of shot coherence to find scene boundaries. A two pass approach is implemented, in the first pass; the coherence of colour features for shots are used to create potential scene boundaries. The second pass involves using motion features, as well as shot lengths to merge scenes and produce a final set of scene boundaries. The ShotWeave system [124], extracts features about each shot using a region based approach, and uses these features to cluster shots and detect scene breaks. The regions used are chosen based on film grammar rules. Troung et al [125] create their own definition of a scene, and implement both edge-based and colour-coherence based approaches to scene boundary detection. Film grammar rules were used to refine the results of both systems, and it was found that the colour-coherence approach yielded better results. Liu et al [126] use hidden Markov models, trained on visual information, to extract a scene-like structure on documentaries.

The use of non-visual features has also been extensively investigated in the literature. In the following section, a brief outline of such techniques is presented. In section 4.1.2 a detailed analysis of such algorithms are presented. Sundaram et al [127] define a computable scene as one which exhibits long term consistency of chromaticity, lighting and ambient sound. A set of audio features are used to determine fundamental changes in the audio, while a coherence-based model is used to locate changes in the visual features. These are then combined to produce a set of scene breaks. Cao et al [128] use the same definition of a scene as in [129], and also use audio to assist in scene boundary detection. However, the assumption is made that most scene boundaries are determined by visual properties. So, in this approach, scene breaks are firstly detected using visual features, and then false boundaries are removed based on the divergence of the audio features. Huang et al [130] note that typically a change in image or motion characteristics is associated with a shot break, while a scene break contains a simultaneous change of image, motion and audio characteristics. Three sets of features are then used to locate breaks in audio, colour and motion and scene and shot breaks are then detected based on these features. Li and Kuo [131] implement a window-based sweep algorithm to locate shot sinks, which contain a pool of shots that are visually similar, and temporally close to each other. The shot sinks are then used to classify scene breaks. The consistency of the audio is then examined to filter the detected scene boundaries.

The approaches presented above are applied for generic content such as documentaries, commercials etc. However a number of algorithms are reported to tackle the domain specific challenges. In this section, an investigation of techniques for particular news content is presented. For the news content scene boundary is commonly referred to story boundaries and hence is used interchangeably. The approach discussed by Merlino and Boykin in [132] uses hidden Markov models to detect the beginning and end of stories within news programs. Audio, visual and text features are used to

generate a set of observable states, and then the hidden Markov model finds the presence of story breaks. Also, O'Hare et al. [133], use a support vector machine based approach to story segmentation. A set of four features (anchorperson shot identification, face detection, motion analysis and shot length analysis) are extracted for each shot, and the support vector machine locates the anchorperson shots which signal the boundary between news stories. One of the tasks in the aforementioned TRECVID workshop is to find story boundaries for news programs. The system developed by Quenot et al [134], is based on detecting changes in the audio associated with a scene break. Changes included a long pause, a change in speaker, as well as the detection of jingles and common phrases commonly spoken by presenters at the end of a news story. The approach implemented by Hoashi et al [135], used a combination of audio, motion, colour and temporal features to assist in story boundary detection. A support vector machine is then used to discriminate between story boundary shots and all other shots. In the summary of the results of the story boundary task in TRECVID, [136] note that a combination of audio, visual, and speech recognition features gives the best overall performance in this task.

Summarisation of Visual Content

With the availability of digital video content, users are increasingly requiring assistance in accessing digital videos [137]. Hence, research into video summarisation helps to meet these needs by creating a condensed version of full length video through the identification of most important and pertinent information within the video stream. Video summarisation techniques produce summaries by analyzing the video content, condensing the content into abbreviated descriptive forms that represent surrogates of the original content embedded with the video [138]. However, a video summary should be different from video trailers where certain contents are intentionally hidden so as to magnify the attraction of a video [139]. Techniques in automatic video summarisation in broad can be categorized into two major approaches: static story board summary [140], [141], [142], [143] and dynamic video skimming [144], [145], [146] and [147]. The former is a collection of static key frames of video shots, while the latter is a shorter version of video composed of a series of selected video clips. Static story board allows non-linear browsing of video content by sacrificing the temporal evolution of a video. Dynamic video skimming, in contrast preserves the time-evolving nature of a video by linearly and continuously browsing certain portions of a video content depending on a given time length. For both approaches the appropriate selection of video segments plays a major role in maximizing the entropy information and perceptual quality of a video summary [139].

To date compared with static story board summary; there is relatively few works that address dynamic video skimming. Nonetheless, due to the advance and popularity of audio-visual capturing tools, effective techniques for dynamic video skimming are highly in demand. A tool that can automatically shorten the original video while preserving most events by highlighting only the important content would be greatly useful especially for news browsing user community. Techniques for dynamic video skimming include applying expectation maximization [148], singular value decomposition (SVD) [144], motion model [149], [150], utility framework [149], [148], attention model [151] and semantic analysis [145], [147]. Visual information has been the primary focus of analysis in creating video summaries, however audio and linguistic information have also been incorporated in order to derive semantic meaning. In [145] audio and motion signals are used to detect emotional dialogues and violent scenes for summarization. However, this approach can only be applied to certain videos, and the resulting summaries may not be useful in revealing the content coverage. In [147] the InfoMedia system was developed to generate the short synopsis of a video. Language understanding techniques are applied with aid of audio and visual features. Nevertheless, this text – driven approach could not generate satisfactory results when speech signals are noisy, which happens frequently in life video recording which is mostly the case in news video too.

Research Focus in Papyrus

In Papyrus, algorithms and tools will be developed for automatic extraction of news stories describing science and technological aspects within the scope of Papyrus. Also, the use of visual descriptors presented in Section 4.1 will be investigated for content structuring problem, along with exploiting underlying news content grammar.

4.2.2. Audio/Speech

This section presents an overview of state-of-the-art approaches for Audio/Speech and Text processing within the scope of Papyrus project. It starts with a description of the common low level descriptors employed for multimedia content analysis in the audio domain (acoustic parameters and MPEG – 7 descriptors). The following section is devoted to audio and speech processing. Beginning with the audio segmentation, which is the first step in audio processing and enables to automatically divide the audio stream in homogeneous segments (e.g. that belongs to a unique category such as speech, music and noise) identifying the nature of the audio signal.

Much emphasis is given here on speaker-based segmentation, which is essential in the content structuring phase and in semantic-based audio stream analysis. We then present an overview of speech recognition systems, which are routinely used to convert spoken documents into a symbolic (phonemic or textual) representation, more amenable to indexing and search.

The final sections are devoted to text analysis, aiming to provide a general overview of existing abilities of large-scale Natural Language Processing (NLP) systems to produce useful features for indexing systems. In particular, we detail techniques used for low-level analysis of texts (morphological analysis, part-of-speech tagging, chunking), and introduce more recent developments in the area of information extraction, which is a first step towards a true "semantic" labelling of textual information. The results of several european projects (ACEMEDIA, BOEMIE and MESH) and networks of excellence (K-SPACE, MUSCLE) have been taken in account to produce this section.

4.2.2.1 Audio descriptors

Low level descriptors

Before introducing content analysis tools for audio and speech, it is noted that an important part of signal processing methods and approaches available in literature are common to many different domains.

In this section we discuss low-level audio features that can be extracted from both uncompressed and compressed audio signals, with applications to audio indexing, analysis, and classification. The term audio here refers to generic sound signals, which include speech, dialog, music, songs, radio broadcast, audio tracks from video programs, noise, and mixtures of any of these signals.

The audio features could be broadly classified into *short-term frame level* and *long-term clip level.* The frame level is defined as a group of neighbouring samples with duration between 10 and 40 ms, so that stationary nature of signals can be assumed. For a feature to reveal the semantic meaning of an audio signal, analysis over a much longer period is necessary, usually from one second to several tens seconds. This interval is called an audio clip and it consists of a sequence of audio frames. The clip boundaries may be the result of audio segmentation such that the frame features within each clip are similar. These two main divisions can be further divided according to their processing domain into *time-domain* features and *frequency-domain* features.

Frame level features

Most of the frame-level features are inherited from traditional speech signal processing. Generally they can be separated into two categories: time-domain features, which are computed from the audio waveforms directly, and frequency-domain features, which are derived from the Fourier transform of samples over a frame (Liu and Wang, 2003) [152].

- ***Time-domain features***
  - *Volume–Energy*: The most widely easy-to-compute frame feature is volume. Volume is a reliable indicator for silence detection, which may help to segment an audio sequence and to determine clip boundaries. The volume on an audio signal depends on the gain value of the recording and digitizing devices. To eliminate the influence of the device, it is possible to normalize the volume of some previous frames.
  - *Zero Cross Rate (ZCR)*: Another time-domain feature is Zero Cross Rate. ZCR is a very useful measure to discern between voiced/unvoiced speech, because typically unvoiced

speech has a low volume, but a high ZCR. With the combination of ZCR and volume, it is possible to classify low volume and unvoiced speech as silent.

- *High Zero Crossing Rate Ratio (HZCRR)*: This derivative time domain feature from ZCR is based on average ZCR values taking benefits from multiple window measurement and highlighting frames with high ZCR.
- *Silence Crossing Rate (SCR)*: it is the number of times that the energy falls below some silence level criterion.

- ***Frequency-domain features***
  - *Pitch or fundamental frequency*: it is obtained by detection and extraction of the fundamental peak from the frequency spectrum. This feature is correlated to the presence of harmonic components in instrumental sounds or in voiced components of speech.
  - *Frequency Centroid (FC)*: This feature is related to the human sensation of the brightness of a sound.
  - *Effective Bandwidth (BW):* Directly related to the frequency centroid and using the previously computed power spectral density it is also possible to calculate its standard deviation that represents a measure of the signal effective bandwidth
  - *Energy Ratio Sub-Band (ERSB)*: To adapt the result to the perceptual property of the human ears it is used the Ratio of Energy in a frequency subband (ERSB). The entire frequency band is divided into four subbands, each consisting of the same number of critical bands, where the critical bands correspond to cochlear filters in the human auditory model.
  - *Mel Frequency Cepstral Coefficients (MFCC)*: One of the most popular set of features used to parameterize the speech is the Mel- Frequency Cepstral Coefficients (MFCC). These are the standard features used in Automatic Speech Recognition (ASR)
  - *Linear-Predictive Coding and Line Spectral Frequencies (LPC and LSF)*: Linear-Predictive Coding, and its derivation Line Spectral Frequencies, have also been widely investigated. The latter are often preferred because they were shown to be strongly related to the vocal tract geometry.
  - *Mean and Variance of the Discrete Wavelet Transform (DWT)*: DWT gives the frequency estimates of a signal at a particular time. The mean and variance of DWT give good discriminating feature vectors.

Clip-Level Features

As described before, frame level features are designed to capture the short-term characteristics of an audio signal; but it is necessary to observe the temporal variation of frame features, if a higher semantic content analysis is performed. This leads to the development of various clip-level features, which characterize how frame-level features change over a clip (Liu and Wang, 2003) [152].

- ***Time-domain features***
  - *Volume-based:* Statistical measures like standard deviation (VSTD) or mean can be computed from the volume feature over a clip.
  - *Energy entropy*: The energy entropy is computed by dividing each audio frame into segments of *K* samples each. The signal energy is computed over each of these segments and normalized by the overall frame energy.
  - *ZCR-based*: Some researchers have reported the usefulness of the standard deviation of the ZCR (ZSTD) to differentiate between TV program categories. According to Saunders [153], statistics of the ZCR can be used to discriminate between speech and music audio segments with high accuracy classification rate.
- ***Frequency-domain features***

- *Pitch-based*: It is not easy to derive the scene content directly from the pitch level of isolated frames, but the dynamics of the pitch envelope of successive frames appear to reveal the scene content more. So it is possible to compute the following clip-level features to capture the variation of pitch: Pitch Standard Deviation (PSTD), Smooth Pitch Ratio (SPR) and Non Pitch Ratio (NPR). SPR is the percentage of frames in a clip that have similar pitch as the previous frames. This feature is used to measure the percentage of voiced or music frames within a clip, since only voiced and music have smooth pitch. On the other hand, NPR is the percentage of frames without pitch and it is used to measure the percentage of unvoiced speech or noise within a clip.
- *Spectrum Flux*: Spectrum flux (SF) is defined as the average variation value of spectrum between the adjacent two frames in a window.

MPEG-7 Audio low level descriptors

The low-level audio Descriptors are of general importance in describing audio. There are seventeen temporal and spectral Descriptors that may be used in a variety of applications. They can be roughly divided into the following groups:

- Basic
- Basic Spectral
- Signal Parameters
- Timbral Temporal
- Timbral Spectral
- Spectral Basis

Additionally, a very simple but useful tool is the MPEG-7 silence Descriptor. The standard audio Descriptors can be seen in Figure 4 and are briefly described below.

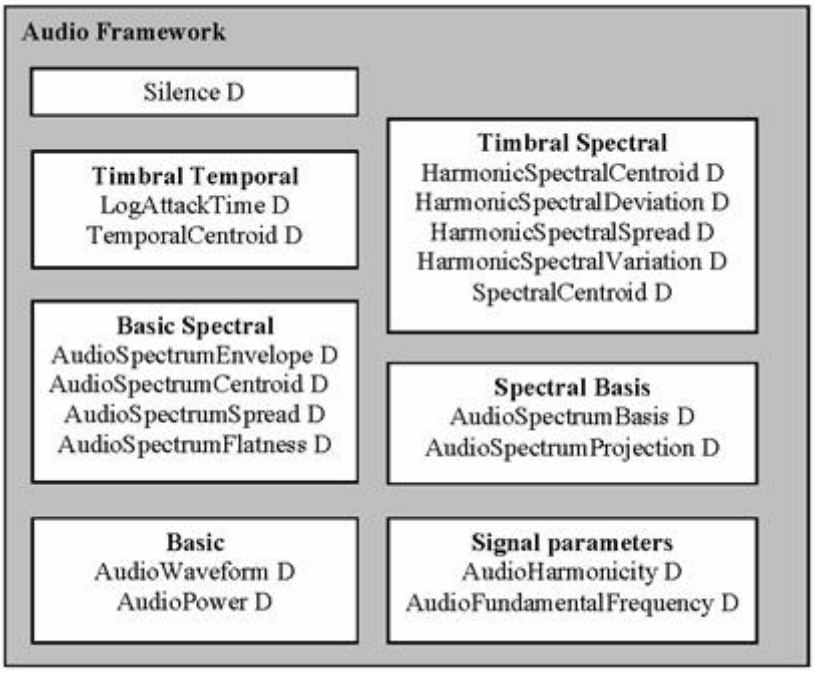

Figure 4-1: Overview of Audio Framework including Descriptors

Basic

The two basic audio Descriptors are temporally sampled scalar values for general use, applicable to all kinds of signals. The AudioWaveform Descriptor describes the audio waveform envelope (minimum and maximum), typically for display purposes. The AudioPower Descriptor describes the temporally-

smoothed instantaneous power, which is useful as a quick summary of a signal, and in conjunction with the power spectrum.

Basic Spectral

The four basic spectral audio Descriptors all share a common basis, all deriving from a single time-frequency analysis of an audio signal. They are all informed by the first Descriptor, the AudioSpectrumEnvelope Descriptor, which is a logarithmic-frequency spectrum, spaced by a power-of-two divisor or multiple of an octave. This AudioSpectrumEnvelope is a vector that describes the short-term power spectrum of an audio signal. It may be used to display a spectrogram, to synthesize a crude "auralization" of the data, or as a general-purpose descriptor for search and comparison.

The AudioSpectrumCentroid Descriptor describes the center of gravity of the log-frequency power spectrum. This Descriptor is an economical description of the shape of the power spectrum, indicating whether the spectral content of a signal is dominated by high or low frequencies.

The AudioSpectrumSpread Descriptor complements the previous Descriptor by describing the second moment of the log-frequency power spectrum, indicating whether the power spectrum is centered near the spectral centroid, or spread out over the spectrum. This may help distinguish between pure-tone and noise-like sounds.

The AudioSpectrumFlatness Descriptor describes the flatness properties of the spectrum of an audio signal for each of a number of frequency bands. When this vector indicates a high deviation from a flat spectral shape for a given band, it may signal the presence of tonal components.

Signal Parameters

The two signal parameter Descriptors apply chiefly to periodic or quasi-periodic signals.The AudioFundamentalFrequency descriptor describes the fundamental frequency of an audio signal. The representation of this descriptor allows for a confidence measure in recognition of the fact that the various extraction methods, commonly called "pitch-tracking," are not perfectly accurate, and in recognition of the fact that there may be sections of a signal (e.g., noise) for which no fundamental frequency may be extracted. The AudioHarmonicity Descriptor represents the harmonicity of a signal, allowing distinction between sounds with a harmonic spectrum (e.g., musical tones or voiced speech [e.g., vowels]), sounds with an inharmonic spectrum (e.g., metallic or bell-like sounds) and sounds with a non-harmonic spectrum (e.g., noise, unvoiced speech [e.g., fricatives like 'f'], or dense mixtures of instruments).

Timbral Temporal

The two timbral temporal Descriptors describe temporal characteristics of segments of sounds, and are especially useful for the description of musical timbre (characteristic tone quality independent of pitch and loudness). Because a single scalar value is used to represent the evolution of a sound or an audio segment in time, these Descriptors are not applicable for use with the Scalable Series. The LogAttackTime Descriptor characterizes the "attack" of a sound, the time it takes for the signal to rise from silence to the maximum amplitude. This feature signifies the difference between a sudden and a smooth sound. The TemporalCentroid Descriptor also characterizes the signal envelope, representing where in time the energy of a signal is focused. This Descriptor may, for example, distinguish between a decaying piano note and a sustained organ note, when the lengths and the attacks of the two notes are identical.

Timbral Spectral

The five timbral spectral Descriptors are spectral features in a linear-frequency space especially applicable to the perception of musical timbre. The SpectralCentroid Descriptor is the power-weighted average of the frequency of the bins in the linear power spectrum. As such, it is very similar to the AudioSpectrumCentroid Descriptor, but specialized for use in distinguishing musical instrument timbres. It is has a high correlation with the perceptual feature of the "sharpness" of a sound.

The four remaining timbral spectral Descriptors operate on the harmonic regularly-spaced components of signals. For this reason, the descriptors are computed in linear-frequency space. The HarmonicSpectralCentroid is the amplitude-weighted mean of the harmonic peaks of the spectrum. It

has a similar semantic to the other centroid Descriptors, but applies only to the harmonic (non-noise) parts of the musical tone. The HarmonicSpectralDeviation Descriptor indicates the spectral deviation of log-amplitude components from a global spectral envelope. The HarmonicSpectralSpread describes the amplitude-weighted standard deviation of the harmonic peaks of the spectrum, normalized by the instantaneous HarmonicSpectralCentroid. The HarmonicSpectralVariation Descriptor is the normalized correlation between the amplitude of the harmonic peaks between two subsequent time-slices of the signal.

Spectral Basis

The two spectral basis Descriptors represent low-dimensional projections of a high-dimensional spectral space to aid compactness and recognition. These descriptors are used primarily with the Sound Classification and Indexing Description Tools, but may be of use with other types of applications as well. The AudioSpectrumBasis Descriptor is a series of (potentially time-varying and/or statistically independent) basis functions that are derived from the singular value decomposition of a normalized power spectrum. The AudioSpectrumProjection Descriptor is used together with the AudioSpectrumBasis Descriptor, and represents low-dimensional features of a spectrum after projection upon a reduced rank basis.

Together, the descriptors may be used to view and to represent compactly the independent subspaces of a spectrogram. Often these independent subspaces (or groups thereof) correlate strongly with different sound sources. Thus one gets more salience and structure out of a spectrogram while using less space.

Silence segment

The silence segment simply attaches the simple semantic of "silence" (i.e. no significant sound) to an Audio Segment. Although it is extremely simple, it is a very effective descriptor. It may be used to aid further segmentation of the audio stream, or as a hint not to process a segment.

High level descriptors

Mpeg-7 spoken content descriptors

Nowadays there is large variety of proposed automatic speech recognition (ASR) systems which would be characterized using a large number of parameters like spoken language, word and phonetic lexicons, quality of the material used to train the acoustic models, parameters of the language models, etc. In order to enable the interoperability between them MPEG-7 Spoken Content high-level description, which is achieved independently of the peculiarities of the recognition engines used to extract spoken content, aims at standardizing the representation of ASR outputs.

There are two distinct elements in the MPEG-7 SpokenContent description:

SpokenContentHeader and SpokenContentLattice. Metadata information such as speaker information, language information, etc. would be saved in SpokenContentHeader and actual decoding produced by an ASR engine would be represented by SpokenContentLattic. Generally the MPEG-7 Spoken Content Description defines a standard output form of lattices delivered by a recognizer.

MPEG-7 SpokenContentHeader descriptor

The *SpokenContentHeader* consists of five types of metadata:

• ***WordLexicon***: generally the vocabulary of a word-based recognizer. A WordLexicon consists of the following elements:

- *phoneticAlphabet*: is the name of an encoding scheme for phonetic symbol and it is needed only if phonetic representations are used;
- *NumOfOriginalEntries*: is the original size of the vocabulary known to the ASR system;
- *Tokens*: each of which stores an entry of the vocabulary and is made up of the *Word* (the label corresponding to the word entry without white space), *representation* (optional, two possible value *orthographic* and *nonorthographic*) and *linguisticUnit*(indicates the type of the

linguistic unit corresponding to the entry, the possible values are: *word*, *syllable*, *morpheme*, *stem*, *affix*, *nonspeech*, *phrase* and *other*).

A header can contain more than one word lexicon;

• ***PhoneLexicon***: each entry of which is an identifier representing a phonetic unit, according to a specific phonetic alphabet. It contains following elements:

- *phoneticAlphabet*: is the name of an encoding scheme for phonetic symbols (*sampa1, ipaSymbol2, ipaNumber3* and *other4*)
- *NumOfOriginalEntries*: is the size of the phonetic lexicon depending on the chosen spoken language and phonetic alphabet;
- *Token*: each of which corresponding to an entry of the phonetic lexicon and must not contain white-space characters.

The same like *WordLexicon* several phone lexicons could be contained in a single header;

• ***ConfusionInfo***: contains confusion statistics computed on a given evaluation collection. The reference transcription and the recognized transcription of a given spoken document are compared by string alignment using dynamic programming to obtain the confusion statistics. Actually this field refer to a description called *ConfusionCount* including following elements:

- *numOfDimensions*: indicates the applied *PhoneLexicon* corresponded dimensionality of the vectors and matrix in the *ConfusionCount* description.
- *Insertion*: is a vector of length *numOfDimension*. It contains the occurrence of a phone inserted in the recognized transcription but not included in reference transcription.
- *Deletion*: is a vector containing the number of a phone presented in reference transcription but deleted in recognized transcription.
- *Substitution*: is a square matrix with a size of *numOfDimension* numOfDimension*. We assume that the rows ***r*** indicated the different phone in reference transcription and the columns ***c*** are the phones in recognized transcription. And the value of (***r, c***) reports that how many times the phone ***r*** in reference transcription is substituted by phone ***c*** in recognized transcription.

• ***SpeakerInfo****:* contains the description of speaker and may be shared by several lattices. The SpeakerInfo has following elements:

- *Person*: is the name (or any identifier) of the speaker
- *SpokenLanguage*: is the spoken language.
- *WordIndex*: consists of a list of words or word n-grams, together with pointers to the place where each words or word n-gram occurs in the lattices concerned.

• ***PhoneIndex:*** consists of a list of phones or phone n-grams, together with pointers to where each phone or phone n-gram occurs in the corresponding lattices.

- *defaultLattice*: is the default lattice for the lattice entries in both the word and phone indexes;
- *wordLexiconRef*: is the reference to the word lexicon used by this speaker;
- *phoneLexiconRef*: is the reference to the phone lexicon used by the speaker;
- *confusionInfoRef*: is a reference to the *ConfusionInfo* description.
- *DescriptionMetadata*: contains the information about the extraction process.
- *Provenance*: indicates the provenance of this decoding. The Provenance could be one of five attributes: *unknown*, *ASR*, *manual*, *keyword* and *parsing*.

The information contained in *SpokenContentHeader* could be shared by several SpokenContentLattice descriptions.

MPEG-7 SpokenContentLattice descriptor

The *SpokenContentLattice* contains the whole information about the decoded lattice. A lattice is described by a series of *block*s. A *block* consists of following elements:

• ***Node**:* is the series of lattice nodes each of which contains timing information.

Following information are described:

- *num:* indicates the number of this node in current block;
- *timOffset:* the time delay to the begin of current block in one- hundredth of a second.
- *speakerInfoRef:* is optional and is a reference to the speaker of current block;
- *WordLink/phoneLink:* contains all links (word/phone) starting from this node and carrying a word/phone hypothesis.

o *Link:* is used to define the lattice links with five attributes: *probability*, *nodeOffset* and *acousticScore*.

• ***MediaTime:*** optionally, it indicates the start time of a block and the duration of this block.

• ***defaultspeakerinfoRef***: is a reference to a SpeakerInfo description.

• ***num***: indicates the range of this block (0-65 535)

• ***audio***: is the measure of the audio quality and has following values:

- *unknown*: which means non information available;
- *speech*: corresponds to clean speech;
- *noise*: corresponds to non-speech;
- *noisySpeech*: indicates that the audio signal is speech but with noise.

Tools for extracting MPEG-7 audio features

Tools and APIs for efficient extraction and update of MPEG-7 XML or binary descriptions have been developed or are under development. This section gives an overview on existing solutions including free-use libraries, APIs and other tools facilitating work with MPEG-7.

Part 6 of the MPEG-7 standard presents the reference software **XM** (MPEG-7 eXperimentation Model[90]). This part contains a reference C++ implementation of its descriptors, the coding schemes and a simple querying application, which was developed for testing and simulation. However, it does not contain a framework, GUI, a documentation of the CBVR (content-based video retrieval) part, optimized descriptor extraction functions and performance optimized algorithms.

Several libraries have been developed with the objective of describing audio content with some descriptors of the MPEG-7 standard, e.g. Java MPEG-7 Audio Encoder[91], Extract Audio Spectrum Envelope[92], SoundSpotter MPEG-7 Audio Software[93]. Implementations are available in Java, C++ or Matlab source under GPL or LGPL.

TU Berlin provides two online demonstrations for the extraction of audio descriptions;

MPEG-7 Audio Analyzer Low Level Descriptors Extractor[94] and Spoken Content Demonstrator[95].

---

[90] MPEG – 7 XM - http://www.lis.ei.tum.de/research/bv/topics/mmdb/e_mpeg7.html

[91] http://mpeg7audioenc.sourceforge.net/

[92] http://sourceforge.net/projects/mpeg7ease

[93] http://www.soundspotter.org/

[94] http://mpeg7lld.nue.tu-berlin.de

[95] http://mpeg7spkc.nue.tu-berlin.de

The Audio Analyzer implements all 17 audio descriptors defined in the standard, thus it computes LLDs of any audio file and receives the chosen LLDs in an MPEG-7 XML file. The LLDs are the basis to create advanced MPEG-7 audio content-based applications. The Spoken Content Demonstrator tool extracts an MPEG-7 SpokenContent description from an input speech signal. The extracted SpokenContent DSs can be used for different types of applications, especially for spoken document retrieval (SDR) applications.

- Mpeg-7 Audio DB
    - Description: A mpeg-7 based audio database, reads mpeg-7 descriptions (see our twin project MPEG7AUDIOENC) and allows useful database like operations.
    - URL: http://sourceforge.net/projects/mpeg7audiodb
    - Type: Standalone application
    - Features: database
    - Input: MPEG-7
    - Output: RDF
    - Operating System:
    - License:
    - DOAP:
- Mpeg-7 Audio ENC
    - Description: The Java MPEG-7 Audio Encoder extracts some descriptors and description schemes of the MPEG-7 standard to describe an audio content (in this case: an audio file)
    - URL: http://mpeg7audioenc.sourceforge.net/
    - Type: Standalone application
    - Features: automatic annotation (feature extraction)
    - Input: WAV, AU, AIFF, MP3
    - Output: MPEG-7
    - Operating System: OS Independent (Written in an interpreted language - Java), OS Portable (Source code to work with many OS platforms)
    - License: GNU Library or Lesser General Public License (LGPL)
    - DOAP:
- MPEG-7 Low Level Audio Descriptors Extractor
    - Description: Extracts 17 Low Level Descriptors (LLDs) defined within the MPEG-7 standard
    - URL: http://mpeg7lld.nue.tu-berlin.de/
    - Type: Web application
    - Features: automatic annotation (feature extraction)
    - Input: WAV, MP3
    - Output: MPEG-7
    - License:
    - DOAP:
- MPEG-7 Feature Extraction Matlab Source

  - Description: The MPEG-7 Audio Reference Software Toolkit. The conformance directory contains examples of extraction for every descriptor from supplied media sources.
  - URLhttp://mpeg7.doc.gold.ac.uk/mirror/index.html
  - Type: MatLab code
  - Features:
  - Input:
  - Output: XML
  - License:
  - DOAP:
- MPEG-7 SpokenContent Description Scheme Extractor
  - Description: A demonstration tool that extracts an MPEG-7 SpokenContent description from an input speech signal.
  - URL: http://mpeg7spkc.nue.tu-berlin.de/
  - Type: Web application
  - Features: automatic annotation (feature extraction)
  - Input: WAV, MP3
  - Output: MPEG-7
  - License:
  - DOAP:
- Transcriber
  - Description: A tool for segmenting, labeling and transcribing speech
  - URL: http://trans.sourceforge.net/en/presentation.php
  - Type: Standalone application
  - Features: manual annotation, automatic segmentation
  - Input: Most standards audio formats (use Snack Sound Toolkin)
  - Output: SGML
  - Operating System: OS Portable (Source code to work with many OS platforms)
  - License: GNU General Public License
  - DOAP:

Research Focus in Papyrus

In the Papyrus project the pertinence of MPEG-7 descriptors to the speech analysis will be analysed and possible extensions with different low level features will be done through the Description Definition Language (DDL), which defines the syntax of the MPEG-7 Description tools and allows the creation of new DSs and Ds (Description Schemes) and the extension and modification of existing DSs.

4.2.2.2 Audio and speech processing

4.2.2.2.1 Audio processing

4.2.2.2.1.1 Audio segmentation

The coarse audio segmentation is the temporal segmentation of an audio signal in segments belonging to different audio categories like speech, music or other general categories. This can be seen as a pre-processing step for more specific audio content analysis. Then, the segments can be analyzed with content-adapted techniques like speech recognition for speech segments or recognition of environmental sounds for non-speech segments.

Kemp et al. [154] and Chen and Gopalakrishnan [155] categorize methods for temporal segmentation of an audio signal into three different approaches:

- Energy-based segmentation for detection of silence with audio energy
- Metric-based segmentation with segment boundaries at maxima of distances between succeeding analysis windows
- Model-based segmentation, which takes advantage of statistical classification methods

Many different approaches have been proposed in the literature for the segmentation of an audio stream into homogeneous parts.

Pfeiffer et al. [156] propose a set of biologically inspired features for automatic audio content analysis. Saunders [152] presents a classifier for speech and music based on zero-crossing rate (ZCR) and short-time energy (STE). Scheirer and Slaney [157] examine 13 different features with statistical classification techniques like GMM and k-nearest-neighbour (k-NN) algorithm for the discrimination of speech and music. An autocorrelation measure, is used in Saraceno [158] for speech and music discrimination based on difference in the fundamental period. Music/speech discrimination with simple threshold decisions considering the quadratic difference between the average ZCR values of frames is performed in Qiao [159] for audio coding. In Minami [160], edge detection is applied on spectrogram and decisions for music segments are determined with a threshold for total edge intensity. Rossignol et al. [161] propose a source segmentation scheme that uses the mean and the variance of spectral flux, spectral centroid, and ZCR as features and is inspired by the work of Scheirer and Slaney. Classification methods (GMM, k-NN, and neural network) are applied for the temporal segmentation task for the two categories speech and music. Carey et al. [162] compare different features together with GMM as classification method for music/speech discrimination using cepstral coefficients, amplitude, pitch, ZCR and the delta values of these features. Their experiments with speech from thirteen languages and music from all over the world indicate that cepstral coefficients and the corresponding delta values give the best performance with GMM classification.

Williams and Ellis [163] propose a speech/non-speech classifier based on four features computed from a phone probability array: mean per-frame entropy, average probability "dynamism", background-label energy ratio, and phone distribution match. Scheirer [157] reports experiment results for speech/music discrimination with audio data. Foote [164] proposes a measure of audio novelty. The examined method identifying natural segment boundaries via analysis of the local self-similarity and is applied to the classification of speech and music. El- Maleh et al. [165] propose the usage of line spectral frequencies (LSP) and ZCR for robust frame-based music/speech discrimination. In experiments with a quadratic Gaussian classifier and a neural network classifier, they have achieved high recognition accuracy. Lu and Hankinson [166] use silence ratio and ZCR basing their experiments on rule-based classification for speech and music audio segments. Ezzaidi et al. [167] propose two systems for music/speech discrimination. The first system uses MFCC as features and GMM for classification. The second system uses delta MFCC values for the computation of distances on frame-level for threshold-based decisions. Harb et al. [168] use a mel-scaled filterbank and piecewise Gaussian modelling combined with a neural network for general audio classification/segmentation. Their experiments include also music/speech discrimination. The usage of Haar-like features derived from spectrogram (FFT, RASTA, Log-FFT) with the AdaBoost algorithm is proposed in [169] for the training of a music/speech classifier. The experiments show in comparison with the results from [157] that better results can be obtained on frame-level with AdaBoost. Muñoz-Expósito et al. [170] introduce the warped LPC-based spectral centroid (WLPC-SC) feature for music/speech discrimination comparing this feature with the known six different timbral features from [173]. The WLPC-SC feature can improve the results of the discrimination task.

Kimber and Wilcox [171] use hidden Markov models (HMMs) for supervised classification and segmentation with cepstral coefficients. For unsupervised classification, they propose a combination of generalized likelihood ratio (Gish distance) with Gaussian mixture models for agglomerative clustering. Experimental results are reported for the following categories (i) five speakers, (ii) seven speaker, laughter, simultaneous talkers, and noise, (iii) instrumental music, female speech, male speech, song with female singer one, song with female singer two, song with male singers, applause, bells, and noisy speech.

Zhang and Kuo [172] first segment an audio signal with short-time energy function, short-time average zero-crossing rate, and short-time fundamental frequency. These segments are classified in the classes silence, environmental sound, music, speech with a rule-based heuristic procedure.

Tzanetakis [173] describes a general methodology for audio segmentation based on peak-detection of the temporal derivative of a distance metric like Mahalonobis distance. The distances are determined on the frame-level. Here, well-known features like MFCC, ZCR, spectral centroid and others are used. Kemp et al. [154] performed experiments for the classification of the audio signal of radio broadcasts into the four categories anchor, field speech, music, and silence. A model-based method and metric-based methods are compared for coarse audio segmentation. The model-based method uses MFCC features and segments the audio signal with the most probable state path of a Viterbi decoded HMM. Each state of the HMM corresponds with one audio category and each class-dependent probability is modelled via diagonal GMMs. The metric-based methods use the three distance metrics: Kullback-Leibler, Gish distance (generalized likelihood ratio, GLR) and entropy loss. These metric-based methods are borrowed approaches from speaker segmentation.

Zhang and Kuo [172] use energy, average ZCR, fundamental frequency, and spectral peak tracks as features and distinguish between the categories speech, music, song, environmental sound, speech with music background, environmental sound with music background, and silence by using a heuristic rule-based procedure.

Lu et al. [175] use high zero-crossing rate ratio (HZCRR), low short-time energy ratio (LSTER), spectrum flux (SF), linear spectral pair (LSP) divergence distance, band periodicity, and noise frame ratio (NFR). They propose a segmentation approach with two stages and hierarchical processing. In the first stage only speech and non-speech segments are identified with the usage of k-nearest-neighbour classifier (k-NN) or LSP vector quantization (LSP-VQ). Then only the non-speech segments are classified further into music, environmental sound, and silence by a rule-based classification scheme.

A coarse classification approach based on support vector machines (SVM) of subsegments and further rule-based smoothing and audio segmentation is proposed in [175]. Different combinations of the features MFCCs, ZCR, STE, brightness, bandwidth, SF, band periodicity, and NFR are examined. The multi-class classification for silence, music, background sound, pure speech, non-pure speech is performed with a bottom-up binary tree classification scheme and Gaussian radial basis kernel SVMs.

Arias et al. [176] use MFCC features comparing GMM with SVM as classification techniques for the classification and segmentation of the categories music, speech, applause and laughter. They found that SVMs obtain slightly better results for fewer training data in comparison with GMM, which achieve slightly better results for a large amount of training data.

Ghaemmaghami [177] proposes an approach based on temporal decomposition (TD), which analyzes only frames selected by an eigen-analysis, to classify segments with GMMs into the categories speech, music, speech-music, and others. For a better computational efficiency, coarse audio classification and segmentation techniques are examined also for audio data in the compressed domain. Patel et al. [178] distinguish between male speech, female speech and music with band energy ratio and pause rate in a rule-based manner with given thresholds. Nakajima et al. [179] classify compressed audio signals into the categories music, speech, and applause with Bayes discriminant function for multivariate Gaussian distribution. Tzanetakis et al. [180] segment an audio signal into music and speech segments by means of features obtained from the MPEG compressed domain. Srinivasan et al. propose [182] an overview about audio and video content analysis with compressed data.

Instead of using energy as an initial confidence measure, Huijbregts et al. [183] uses the output of a system that is only trained on silence and speech. The training is done on a small amount of Dutch broadcast news training data (three and a half hours of speech and half an hour of silence from 200 male and 200 female speakers). The hypothesis is that, because energy is not used as a feature and most non-speech data will fit the more general silence model better than the speech model, most non speech will be classified as silence. After the initial segmentation the data classified as silence is used to train a new silence model and a sound model. The silence model is trained on data with low energy levels and the sound model on data with high energy levels. After a number of training iterations, the speech model is also re-trained. The result is an HMM based SAD system with three models (speech, non-speech and silence) that are trained solely on the data under evaluation.

Research Focus in Papyrus

Audio segmentation is necessary in our context in order to filter out the audio that do not contain speech. Processing non-speech portions of the videos would in fact introduce noise in the transcripts due to assigning word labels to non-speech fragments and reduce speech recognition accuracy. Different methods will be investigated in order to achieve the best results in the speaker segmentation and the speech recognition task.

4.2.2.2.2 Speech processing

With the ever-increasing number of television (TV) channels and radio stations, many hours of TV and radio broadcasts are collected every year by national heritage institutions and private companies. Apart from the architectural problems underlying the design of databases for storing these data, another crucial problem is information retrieval. In audio data files, information retrieval is normally performed by indexing the audio databases, associating each audio document with a file describing its structure in terms of retrieval keys [184] According to Gartner [185] technologies for audio mining and speech analytics are expected to reach maturity in 3-8 years although audio search is already applied in nich sector like broadcasters.

To perform full indexing, an essential initial step is to determine which speaker is speaking at a given time. This process is known as "speaker segmentation" of the audio files.

Speaker segmentation is the task of dividing an input speech signal into homogenous regions containing the speech of exactly one speaker. For a given speech/audio stream, speaker segmentation/change detection systems find the times when there is a change of speaker in the audio. On a more general level, acoustic change detection aims at finding the times when there is a change in the acoustics in the recording, which includes speech/non-speech, music/speech and others. Acoustic change detection can detect boundaries within a speaker turn when the background conditions change.

Speaker clustering techniques and algorithms agglutinate together segments that belong to the same speaker. This does not entail whether such segments come from the same acoustic file or different ones. It also does not say anything about how acoustically homogeneous segments within a single file are obtained.

The term speaker diarization refers to the systems that perform a speaker segmentation of the input signal and then a speaker clustering of the created segments into homogeneous groups (or some hybrid mechanism doing both at the same time), all within the same file or input stream.

4.2.2.2.2.1 Speaker segmentation and clustering

Acoustic parameters

Speaker segmentation falls into the category of the speaker-based processing techniques. Features extracted from the acoustic signal are intended to convey information about the speakers in the conversations in order to enable the systems to separate them optimally.

Likewise speaker and speech recognition systems, well used parametrization features in speaker segmentation, clustering and diarization are Mel Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), Perceptual Linear Predictive (PLP), Linear Predictive Coding (LPC) and others.

Although the aforesaid parametrization techniques [186] yield a good performance in current speaker diarization and recognition systems, they are usually not focused on representing the information relevant to distinguishing between speakers and to isolate such information from other interfering sources (like non-stationary noises, background music and others). Nevertheless speaker recognition and diarization systems use MFCC parameters with a higher number of coefficients as it is known that the higher coefficients do incorporate speaker information.

Anguera [186] gives an overview of alternative parameters focusing on the speaker characteristics and/or particular conditions of the tasks that they are applied to, all within the speaker-based area, which can constitute an advantage if used alone or in conjunction with the most common parametrization techniques. Although the use of these parameters is still not general, these should constitute the tip of the iceberg of parameters exploiting speaker information to come.

Yamaguchi, Yamashita and Matsunaga [187] propose a speaker segmentation system using energy, pitch frequency, peak-frequency centroid and peak-frequency bandwidth, and adds three new features: temporal feature stability of the power spectra, spectral shape and white noise similarities; all three related to the cross correlation of the power spectrum of the signal.

In order to avoid the influence of background noises and other non-speaker related events, in Pelecanos and Sridharan [188] and more recently in Ouellet, Boulianne and Kenny [189], feature warping techniques are proposed to change the shape of the p.d.f. of the features to a Gaussian shape prior to their modeling. They have been applied with success in Sinha, Tranter, Gales and Woodland [190] and Zhu, Barras, Lamel and Gauvain [191] for speaker diarization in broadcast news and meetings respectively.

In Moh, Nguyen and Junqua [192], Tsai, Cheng and Wang [193] and Tsai, Cheng, Chao and Wang [194] speaker diarization systems are proposed by constructing a speaker space from the data and projecting the feature vectors in it prior to the clustering step. Similarly, Collet, Charlet and Bimbot [195] proposes the technique of anchor modeling (introduced in Sturim, Reynolds, Singer and J.P.Campbell [196]) where acoustic frames are projected into an anchor model space (previously defined from outside data) and performs speaker tracking with the resulting parameter vectors. They show that it improves robustness against outside interfering signals and they claim it to be domain independent.

Chan, Lee, Zheng and hua Ouyang [197] propose the use of vocal source features for the task of speaker segmentation using a system based on Delacourt and Wellekens [198]. Also in Lu and Zhang [200] a real-time 2-step algorithm is proposed by doing a bayesian fusion of LSP, MFCC and pitch features.

Finally, in Kotti, Benetos and Kotropoulos [201] is shown the use of features deriving from mpeg-7 like AudioWaveformEnvelope and the AudioSpecrtumCentroid after investigating several ones.

Methods for speaker segmentation

Different approaches for segmentation of audio records have been proposed in recent years. They can be classified into three main groups:

- Metric based segmentation
- Non metric based segmentation
- Other techniques

Metric based segmentation

Metric based segmentation is probably the most used technique up to date. It relies on the computation of a distance between two acoustic segments to determine whether they belong to the same speaker or to different speakers, and therefore whether there exists a speaker change point in the audio at the point being analyzed. The two acoustic segments are usually next to each other (in overlap or not) and the change-point considered is between them. Most of the distances used for acoustic change detection can also be applied to speaker clustering in order to compare the suitability that two speaker clusters belong to the same speaker.

Let's consider two audio segments (i, j) of parameterized acoustic vectors X_i and X_j of lengths N_i and N_j respectively, and with mean and variance values μ_i, σ_i and μ_j, σ_j. Each one of these segments is modeled using Gaussian processes $M_i(\mu_i, \sigma_i)$ and $M_j(\mu_j, \sigma_j)$, which can be a single Gaussian or a Gaussian Mixture Model (GMM). On the other hand, let's consider the agglomerate of both segments into X, with mean and variance μ, σ and the corresponding Gaussian process $M(\mu, \sigma)$.

In general, there are two different kinds of distances that can be defined between any pair of such audio segments. The first kind compares the sufficient statistics from the two acoustic sets of data without considering any particular model applied to the data, which from now on will be called statistics-based distances. These are normally very quick to compute and give good performances if Ni and Nj are big enough to robustly compute the data statistics and the data being modelled can be well defined using a single mean and variance.

A second group of distances are based on the evaluation of the likelihood of the data according to models representing it. These distances are slower to compute (as models need to be trained and evaluated) but can achieve better results than the statistics-based as bigger models can be used to fit more complex data. These will be referred as likelihood-based techniques.

- ***Distance-based methods***

This approach consists of measuring the dissimilarity between two adjacent windows of (parameterized) audio data. Depending on the degree of dissimilarity, the system locates a change mark at the point at which the dissimilarity is maximized. Several dissimilarity measures have been proposed in the literature, such as the generalized likelihood ratio [202], [203], the Kullback-Leibler distance, or the Bayesian information criterion (BIC) [155], [204]. The main drawback is the presence of a threshold which has to be tuned for each kind of audio database. The DISTBIC algorithm [155] can be considered a two-pass segmentation method belonging to this group. In the first pass, the generalized likelihood measure is used for determining the approximate situation of the segment boundaries; while in the second pass, these changing points are refined by applying the BIC criterion. This algorithm is able to detect any number of speakers, but only detects changes between them, without identifying which one is involved in the change.

- ***Clustering-based methods***

Speaker segmentation of audio data files can be carried out by using a group average hierarchical agglomerative clustering algorithm as proposed in [205]. This technique consists of dividing the audio data into a certain number of segments (clusters) and iteratively merging two clusters according to a predetermined metric. The clustering algorithm uses the information about the number of speakers as the stopping criterion.

In the initialization stage of the clustering algorithm, the data are divided into segments of equal length and this determines the resolution of the segmentation procedure.

The clustering-based approach not always yields good results in it, but can be used to provide initial labelled data in model-based unsupervised segmentation, as in Salcedo [184].

Non Metric-Based Segmentation

- ***Energy-based methods***

These techniques detect speaker changes hypothesizing that most changes between speakers will be through a silence segment. These have been traditionally implemented towards using the segments for speech recognition, as it is very important to obtain clean speaker changes without cutting any words in half. Systems falling into this category are energy-based and decoder-based systems.

The energy-based systems use an energy detector to find the points where it is most probable to exist a speaker change. The detector normally obtains a curve with minimum/maximum points in potential silences. A threshold is usually used to determine them (Kemp et al. [206], Wactlar et al [207], Nishida and Kawahara [208]). In Siu, Yu and Gish [209] the MAD (Mean absolute deviation statistic),

which measures the variability in energy within segments, is used instead in order to find the silence points.

In contrast, decoder-guided segmenters run a full recognition system and obtain the change points from the detected silence locations (Kubala, Jin, Matsoukas, Gnuyen, Schwartz and Machoul [210] , Woodland, Gales, Pye and Young [211], Lopez and Ellis [212], Liu and Kubala [213], Wegmann, Scattone, Carp, Gillick, Roth and Yamron [214]) they normally constrain the minimum duration of the silence segments to reduce false alarms. Some of these systems use extra information from the decoder, such as gender labels (Tranter and Reynolds [215]) or wide/narrow band plus music detectors (Hain, Johnson, Turek, Woodland and Young [216]).

The output has normally been used as input to recognition systems, but not for indexing or Diarization as there is not a clear relationship between the existence of a silence in a recording and a change of speaker. In such systems they sometimes take these points as hypothetic speaker change points, and then using other techniques define which of them actually mark a change of speaker and which do not.

- ***Model-based segmentation***

In this case, a statistical model (for example an HMM [205],[218],[219]) is trained for a set of predefined acoustic classes (speech, speaker, background noise, music, telephone speech, etc.). For segmentation purposes, each frame (or various frames) of the audio stream is classified using a maximum likelihood criterion, and the segment boundaries are located at the temporal point where a change of acoustic class occurs.

The main disadvantages of this method include the need to predefine the number and nature of the acoustic classes and the large quantity of labelled data needed for building the different acoustic models in a supervised manner. This last drawback can be tackled using an initial segmentation of the database, as has been shown in [205].

Segmentation Using Other Techniques

Some speaker segmentation techniques don't fit to any of the previous categories.

Vescovi et al. [220] and Zdansky and Nouza [221] propose dynamic programming to find the speaker change points. In Zdansky and Nouza [221] BIC is used as marginal likelihood, solving the system via ML where all possible number of change points is considered. Vescovi et al. [220] also use BIC exploring possible computation reduction techniques.

In Pwint and Sattar [222] a genetic algorithm is proposed where the number of segments is estimated via the Walsh basis functions and the location of change points is found using a multi-population genetic procedure. In Salcedo-Sanz et al. [184] a genetic algorithm is proposed, for a two speaker segmentation problem, which encodes possible segmentations. The GA fitness function is provided by a measure of the Mutual Information between the samples of audio and the individuals of the GA. The performance of the GA is improved by introducing a more compact encoding of the GA, by means of a compaction factor.

In Lathoud, McCowan and Odobez [223] segmentation is based on the location estimation of the speakers by using a multiple-microphone setting. The difference between two locations is used as a feature and tracking techniques are employed to estimate the change points of possibly moving speakers.

In [224] speaker turn detection and speaker clustering are based on one-class SVM dissimilarity measure. Being insensitive to the dimension of the acoustic features, their algorithm allows the use of more informative, larger dimensional acoustic features.

Tools for speaker segmentation and clustering

MClust

http://www-lium.univ-lemans.fr/tools/index.php?option=com_content&task=blogcategory&id=29&Itemid=56

mClust is a software package dedicated to speaker diarization (i.e. speaker segmentation and clustering). Most of the tools in the package consider segmentation as input and generate a new segmentation as output. The provided tools allows the user to perform BIC hierarchical clustering, Viterbi decoding using GMM models trained by EM or MAP, and CLR hierarchical clustering using GMM (clustering based over automatic speaker recognition methods).

Alize/Mistral

http://mistral.univ-avignon.fr/en/index.html

A module inside the Mistral package is devoted to speaker segmentation: LIA_SpkSeg. Mistral is an open source software for speaker recognition developed at the University of Avignon.

Audioseg

http://gforge.inria.fr/projects/audioseg

AudioSeg provides a range of generic modules and tools in C language for audio analysis, segmentation and classification (including speaker segmentation and verification). It is distributed freely under the GPL agreement. The AudioSeg toolkit is used by several labs, mostly in France. The toolkit implements standard reference algorithms such as energy-based silence detection, BIC segmentation and clustering as well as GMM/HMM classification. It has been benchmarked in the NIST Speaker Recognition Evaluation (in 2004 and 2005) and in the French ESTER evaluation (2005). It was made available to all ESTER participants.

Research Focus in Papyrus

We will attempt to provide an extension of the most common parametrization techniques to incorporate acoustic parameters that focus on the speaker characteristic.

Different methods of segmentation and combinations of them will be used for the speaker segmentation task in order to provide the most reliable results to the content structuring, without loosing generality. For this purpose we will avoid the implementation of algorithms that imply thresholds tuning, as their performance might decrease as database changes (when data come from a new archive).

Results will be evaluated on the basis of international standards (NIST) and on samples of the material provided by the project.

Evaluation will be based on the well known measures of FAR (False Alarm Rate), MDR (Missed Detections Rate), Precision, Recall and F and on other more specific indicators such as the percentage of missed detection of short speaker turns.

4.2.2.2.2.2 Automatic Speech Recognition

The finest level of feature extraction from audio stream is represented by automatic speech recognition. The abbreviation ASR (Automatic Speech Recognition) refers to multiple cross-knowledge and application domains (like acoustic, phonetics, linguistic and lexical domains) where many different tools are used jointly forming a complex infrastructure.

Progress in the field has been driven by standardised metrics, corpora and benchmark testing through NIST since the mid1980s, with systems developed for evermore challenging tasks or 'speech domains': developing from the domain of single person dictation systems to today's research into systems for the meetings and lectures domain. A brief history of speech (and speaker) recognition research can be found in Furui [225].

ASR consists of two major phases: the acoustic feature extraction phase and the decoding phase. The acoustic features (MFCCs or PLP) are extracted from a time frequency analysis of the input speech signal during the feature extraction phase. And the decoding phase aims at determining the most probable sequence of symbols, knowing the acoustic observations, with the help of well-trained acoustic models, language models and pronunciation dictionaries.

Most current ASR systems are based on the hidden Markov model (HMMs) paradigm, which can model any kind of speech units (words, phone, syllable etc.) and allows designing systems with diverse degree of complexity [226].

Due to the increase of the computation power, ASR systems have reached a sufficient level of performance, which makes them useable in many commercial products, from interactive vocal services to dictation programs.

Depending on the task, ASR systems can be classified in following categories [226]:

- connected word recognition which can be used to build a connected digit recognizer;
- large-vocabulary continuous speech recognition (LVCSR) which can be used to build a dictation system, an audio indexing system or a dialogue system;
- automatic phonetic transcription;
- key-words spotting.

Word-error-rate (WER), which is defined as the (the number of insertion errors + the number of deletion errors + the number of substitutions) divided by the number of words in the reference transcription, is often used to evaluate speech-recognition systems.

Over the past decades great strides have been made in speech technology. But robust speech recognition is still far from a solved problem. Depending on the task and test conditions there is a huge performance difference between current systems (WER from 1% to 40%).

Recent developments [225] in the area of speech recognition during the 2000s are:

- Development of speech-to-text (automatic transcription) technology with the aim of achieving substantially richer and much more accurate output than before. The tasks include detection of sentence boundaries, fillers, and disfluencies (DARPA EARS project).
- Spontaneous speech recognition: although read speech and similar types of speech, e.g. news broadcasts reading a text, can be recognized with accuracy higher than 95% using state-of-the-art speech recognition technology, recognition accuracy drastically decreases for spontaneous speech. Broadening the application of speech recognition depends crucially on raising recognition performance for spontaneous speech.
- Robust speech recognition: To further increase the robustness of speech recognition systems, especially for spontaneous speech, utterance verification and confidence measures are being intensively investigated. In order to build acoustic models more sophisticated than conventional HMMs, the dynamic Bayesian network has recently been investigated.
- Multimodal speech recognition: Humans use multimodal communication when they speak to each other. Studies in speech intelligibility have shown that having both visual and audio information increases the rate of successful transfer of information, especially when the message is complex or when communication takes place in a noisy environment. The use of the visual face information, particularly lip information, in speech recognition has been investigated, and results show that using both types of information gives better recognition performances than using only the audio or only the visual information, particularly in noisy environment.

The potential of ASR-based indexing has been demonstrated most successfully in the broadcast news domain [183]. Spoken document retrieval in the American-English broadcast news (BN) domain was even declared 'a solved problem' based on the results of the TREC Spoken Document Retrieval (SDR) track in 1999 [183]. Partly because collecting data to train recognition models for the BN domain is relatively easy, word-error-rates (WER) below 10% are no longer exceptional, and ASR transcripts for BN content approximate the quality of manual transcripts, at least for several languages.

Systems have also been developed for some domains in many other major European languages e.g. the LIMSI-CRNS spoken language processing group has developed broadcast news transcription

systems for French, German, Portuguese and Spanish in addition to English, Mandarin and Arabic (Gauvain & Lamel, 2003) [227].

Research Focus in Papyrus

In the Papyrus project the use of external complementary sources (such as Wikipedia) and internal (manual transcripts and textual news from AFP and DW) for the creation of specific language models will be investigated. Results will be evaluated on test material provided by the project on the basis of performance measures such as WER, Recall, Precision and F-score.

4.2.2.2.2.3 Tool for Automatic Speech Recognition

Open Source/Freeware

- Sphinx (http://cmusphinx.sourceforge.net/html/cmusphinx.php): Sphinx is a speaker-independent large vocabulary continuous speech recognizer released under a BSD style license. It is also a collection of open source tools and resources that allows researchers and developers to build speech recognition systems. The three systems (Sphinx2, Sphinx3, Sphinx4) developed by Sphinx Group at Carnegie Mellon University have been made available in order to stimulate the creation of speech tools and applications and to advance the state of the art both in speech recognition and related areas. Sphinx-2 is meant as a real-time engine and is regarded as appropriate for systems that require short response times. Sphinx-3 is slower but potentially more and Sphinx-4 is a Java implementation. The Sphinx group also makes available acoustic and language models for those wishing to skip aspects of training and data preparation, and tools for acoustic and language model production.
- Sonic (http://cslr.colorado.edu/beginweb/speech_recognition/sonic_main.html): SONIC is a toolkit for enabling research and development of new algorithms for continuous speech recognition. Since March of 2001, SONIC has been used as our test bed for research activities that include speech recognition as core components at the Center for Spoken Language Research. SONIC is an end-to-end solution which can allow one to design, train, and test on many state-of-the-art speech recognition tasks. The recognizer can run in batch-mode (process many audio files sequentially) or in live-mode (speak into the microphone and see output in real-time). It is freeware for research purposes but not open source.
- Julius (http://julius.sourceforge.jp/en_index.php?q=en/index-en.html): Julius is an open source speech recognition engine. Julius is a high-performance, two-pass large vocabulary continuous speech recognition (LVCSR) decoder software for speech-related researchers and developers. Based on word 3-gram and context-dependent HMM, it can perform almost real-time decoding on most current PCs in 20k word dictation task. Major search techniques are fully incorporated. It is also modularized carefully to be independent from model structures, and various HMM types are supported such as shared-state triphones and tied-mixture models, with any number of mixtures, states, or phones. Standard formats are adopted to cope with other free modeling toolkit. The main platform is Linux and other Unix workstations, and also works on Windows. Julius is open source and distributed with a revised BSD style license. Although Julius is only distributed with Japanese models, the VoxForge project is working on creating English acoustic models for use with the Julius Speech Recognition Engine.
- HTK (http://htk.eng.cam.ac.uk/): The Hidden Markov Model Toolkit (HTK) was originally developed at the Speech Vision and Robotics Group (now the Machine Intelligence Laboratory) of the Cambridge University Engineering Department and contains a set of library modules and tools available in C source form used primarily for speech recognition research. It is anticipated that HTK be used at least for phone level transcription and also for optional contextual strategies HTK also contains editing and re-estimation tools.

- Shout (http://wwwhome.cs.utwente.nl/~huijbreg/shout/index.html): Shout is a software package developed at the University of Twente. It is available under request. It contains also tools for speech detection and speaker segmentation.
- Sirocco (http://perso.enst.fr/~sirocco/index-en.html): Sirocco is a project aiming at making available the source code of a large vocabulary speech recognition system based on continuous density Hidden Markov Models. The current package is the result of this effort to distribute such a code. The current distribution contains the source codes of the system (which means that no resources such as acoustic models, lexicons, ... are included in this distribution).

Commercial software

- Nuance: Nuance Communications Inc. produces commercial solutions for speech related technologies including speech recognition, speaker recognition and speech synthesis. Nuance speech recognition features a distributed client-server architecture enabling separation of light client processing from CPU-intensive server processing. Alternatively, for small configuration or for prototyping, the client and server side applications can run in a single-tier configuration. Primarily developed for telephony-based applications, Nuance speech recognition software accepts speaker-independent, continuous speech and supports very large vocabularies. Included is a "template matching" natural language capability for identifying the meaning of speech. A toolkit is available for use in developing a wide variety of speech recognition applications.
- Lumenvox: The LumenVox Speech Engine is an accurate, standards-based speech recognizer that supports multiple languages and can perform speech recognition on audio data from any audio source. On Linux or Windows, the speaker and hardware independent Speech Engine powers speech solutions and platforms deployed in Enterprise and SMB environments worldwide. It also provides speech application developers with an efficient development and runtime platform, allowing for dynamic language, grammar, audio format, and logging capabilities to customize every step of their application. Grammars are entered as a simple list of words or pronunciations, or in the industry standard Speech Recognition Grammar Specification (SRGS).
- BBN: BBN's Byblos iis an automatically trainable system that utilizes probabilistic hidden Markov models, and it continues to represent the state of the art in large-vocabulary, speaker-independent speech recognition. BBN is working also in the filed of speaker identification.
- Microsoft: MSS is a platform for supported integrated speech services including telephony (voice-only) and multimodal (voice/visual) applications. MSS combines Web technologies, speech-processing services, and telephony capabilities within a single system performing speech recognition and speech synthesis for applications that can be accessed by telephone, cell phone, Pocket PC, Tablet PC and other devices. MSS includes the Microsoft Speech Recognition Engine but also supports third-party options like the ScanSoft/SpeechWorks OpenSpeech Recognizer.
- Loquendo: Loquendo ASR is the next-generation speech recognition technology for speech-enabled applications. It is speaker-independent and reliably recognizes large-scale vocabulary continuous speech, even in the noisiest environments such as wireless.
- Virage/Autonomy: Virage SoftSound delivers audio processing applications to enable live or recorded speech to be manipulated, edited, searched and hyperlinked as easily as text. This is achieved with a wide range of speech processing technologies from audio segmentation and identification through to automatic speech recognition and understanding. Virage SoftSound was founded in 1995 and is backed by over ten years of research from Cambridge University. In May 2000 Virage SoftSound received substantial investment from Autonomy and is now a leading provider and developer of speech technology. Continual research and development allow Virage SoftSound to provide state-of-the-art speech processing technology to the Autonomy group of companies.

Other companies in the field on audio mining and speech analytics [185] are: CallMiner, Nexidia, Nice Systems, StreamSage, Telisma, Utopy, Verint Systems, Witness Systems.

4.2.2.3 Linguistic Analysis of Speech Transcripts

The automatic analysis of video/audio/images material is mostly resulting in so-called "low-level" content features (colour, texture and shape) . Comparing to the way humans perceive and access multimedia content, we notice that there is a "semantic gap" in the field of automatic content detection (and indexing) in video processing.

The integration of linguistic and semantic information in the analysis of audio-video data can help to reduce this "semantic gap".

Text associated to a video is available in different forms:

- the OCR processing of texts found in images (including subtitles),
- the information that can be extracted from the transcripts of speech extracted from the audio stream,
- the analysis of accompanying or related text.

In this section we concentrate on this second case.

This kind of analysis is intended to help providing semantic metadata in several tasks, most of them related to the objective of offering high-level (semantic) descriptors of the content of videos. It is expected that the information that can be extracted from the associated transcripts can improve the semantic indexing of video material.

We have to be aware that the generation of transcripts by ASR is error prone, and the linguistic analysis of transcripts has to take this fact into consideration if one wants to avoid coming along with the extraction of erroneous information from transcripts.

When approaches from Natural Language Processing (NLP) on text are applied to errorful speech recognition output, performance significantly degrades when compared to performance on human transcripts [229] below.

Current standard output exhibits several characteristics that pose serious difficulties, both for human comprehension and for subsequent automated processing [228] below.

• **Absence of punctuation**: Most crucially, this means that sentence boundaries are not indicated. Neither is speaker change, or quotation; and phrase boundaries are similarly unmarked.

• **Absence of capitalization**: For named-entity extraction in text, capitalization is by far the best single cue for identifying named entities — generally, proper nouns. ASR output only capitalizes words that are capitalized in the ASR system vocabulary. Thus, out-of-vocabulary proper nouns words and multi-word named entities will generally be incorrectly uncapitalized in ASR output. Exacerbating this problem, named entity will be disproportionately out-of-vocabulary for any standard dictionary.

• **Speaker disfluencies**: Spoken language exhibits unique characteristics not found in written language, such as mid-utterance corrections, frequent use of phrases that are not complete sentences, filler words ("umm", "ah", "err . . ."), and the abrupt start and stop of phrases. Even perfectly transcribed conversation, lacking the prosodic and other cues of live speech, will be very difficult to understand.

• **ASR errors**: Current standard ASR typically exhibits word error rates of 10-50% for large vocabulary, speaker-independent ASR. Speaker-dependent ASR (training an ASR for one particular speaker) and limited-vocabulary ASR (as in systems for navigating voice mail) can achieve much better results, but in clearly much more limited domains. For the needs of intelligence analysis, as well as many commercial technologies, the necessary task is large-vocabulary speaker-independent ASR.

So there is a need to apply very robust language analysis methods in this context. The tools most frequently applied in this context are statistical Part-of-Speech (POS) taggers, Named Entity Recognizers (NER) and a shallow chunkers. The chunker proposes a grouping of words in larger units

(typically called phrases), giving to the textual transcript a first and shallow linguistic structure, and going thus beyond keyword approaches for extracting indexes from speech transcripts. Techniques to insert punctuations to an ASR transcripts can be applied (Hillard, 2008) [229] below as a preprocessing step.

In the BOEMIE http://www.boemie.org/ (Bootstrapping Ontology Evolution with Multimedia Information Extraction) european project (FP6-027538), a deliverables "D2.3 Semantics extraction from non-visual content tools: state-of-the-art report" has been produced [230] below.

From this study, it seems that the majority of approaches follow similar processing steps (from language pre-processing to named entity recognition, co-reference resolution and template filling). They are also rather application-oriented ones, although they have functionalities that enable domain adaptation, since the goal is to develop IE systems that get better evaluation results in specific application areas. This is probably justified by the influence of the IE evaluation conferences during the 90's (Message Understanding Conferences: MUCs), which established a decomposition of IE into standard processing steps, as well as by the influence of machine learning techniques which facilitated experimenting and porting to new application areas.

However, it is good to see that the pre-processing tasks, such as named entity recognition, are now more mature, establishing thus an infrastructure upon which new techniques with stronger involvement of knowledge models (i.e. ontologies) can be exploited. Furthermore, the use of general purpose language processing tools and IE engines in more systems along with the re-use of existing resources (WordNet is employed in several systems) and ontologies, shows that although the goal is IE systems that perform well in specific application areas, the trend is towards the development of IE infrastructures that enable the development of application specific IE systems.

Within the MUSCLE Network of Excellence http://www.muscle-noe.org/ on multimedia understanding, data mining and machine learning researchers have developed a range of tools for text analysis, text annotation, Natural Language Processing text classification and semantic indexing.

Morphological analysis

Morphological analysis is concerned with the inflectional, derivational, and compounding processes in word formation in order to determine properties such as stem and inflectional information. Together with part-of-speech (PoS) information this process delivers the *morpho-syntactic* properties of a word. As a crucial pre-processing step, morphological analysis is used in virtually all fields of NLP and in applications such as information retrieval. Some well-known systems are PC-KIMMO [231], GERTWOL [232], Morphix [233], Mmorph [234], ChaSen [235], the Xerox MLTT system [236], and MULTEXT [237]. Morphological analysis gives information on the stem of a word, possible parts-of-speech (*substantive, adjective, verb, etc.*), its inflectional properties (information on gender: *male, female, neuter; number: plural, singular; case: nominative, accusative, dative, etc.*) and a possible compound analysis (specifically for languages such as German and Dutch). It is based on the existence of an available lexicon for the language under consideration. Since words are very often highly ambiguous with respect to PoS and inflection, some disambiguation steps are needed. Morphological disambiguation interacts with PoS tagging as well as with chunking, which is putting words together in one fragment and thus provides an indirect morphological and PoS disambiguation.

Part of speech tagging

Part-of-speech tagging (POS tagging or POST), also called grammatical tagging, is the process of marking up the words in a text as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e., relationship with adjacent and related words in a phrase, sentence, or paragraph. Once performed by hand, POS tagging is now done in the context of computational linguistics, using algorithms which associate discrete terms, as well as hidden parts of speech, in accordance with a set of descriptive tags.

Currently available tools for PoS tagging are based either on rule-based or stochastic methods to disambiguate and to tag unknown words.

Rule-based approaches use hand crafted or automatically extracted rules that use contextual information to assign tags to unknown or ambiguous words (see for instance: [238], [239], [240]).

For example, such a rule could describe the fact that if a word is preceded by a determiner and followed by a noun, is should be tagged as an adjective. In addition, many rule-based systems use morphological information to aid in the disambiguation process. For instance, a morphological rule could state that a word, which is preceded by a verb and ends on -ing should be tagged as a verb.

Stochastic PoS taggers are based on statistical models, incorporating frequency or probability (see for instance: [241],[242],[243]). A simple stochastic tagger disambiguates words solely on the probability that a word occurs with a particular tag in a given training set. In other words, the most frequent tag in the training set will be the one assigned to an ambiguous instance of that word. A more advanced alternative to this is to calculate the probability of a given sequence of tags (a so-called *n-gram*), i.e. the probability that a tag occurs with the *n* previous tags. The most common algorithm for implementing an n-gram approach is *Viterbi*, a breadth-first search algorithm [244].

Transformation-Based Tagging, sometimes called Brill tagging, is an instance of the Transformation-Based Learning (TBL) approach to machine learning [245], and draws inspiration from both the rule-based and stochastic taggers. Like the rule based taggers, TBL is based on rules that specify what tags should be assigned to what words. But like the stochastic taggers, TBL is a machine learning technique, in which rules are automatically induced from the data. Like some but not all of the HMM taggers, TBL is a supervised learning technique; it assumes a pre-tagged training corpus.

The various part-of-speech tagging algorithms we have described can also be combined. The most common approach to tagger combination is to run multiple taggers in parallel on the same sentence, and then combine their output, either by voting or by training another classifier to choose which tagger to trust in a given context.

Chunk and light parsing

Shallow parsing (also chunking, "light parsing") is the analysis of a sentence which identifies the constituents (noun phrases, verbs,...), but does not specify their internal structure, nor their role in the main sentence. Abney pioneered the idea of parsing by chunks supported by psychological evidence of human parser [246], where chunks are taken to be some non-recursive cores of major phrases. He also tried partial parsing of unrestricted text with finite-state cascades [247] in a knowledge-intensive way. The problem of chunking is further reformulated as a task similar to POS tagging [248], i.e., by adopting a tag set of {B, I, O} combined with chunk type of phrases for those non-overlapping chunks, where:

B: initial word of a chunk

I: non-initial word of a chunk

O: word outside of any chunk

Therefore many learning approaches to POS tagging become directly available for chunking (see, e.g., [249],[250]). Syntactic chunking (partial parsing) of unrestricted written text have become a relatively well-defined and well studied since the introduction of CoNLL 2000 shared task [251]. But the chunking of spontaneous spoken language has received less attention (except [250]) than that of written language though spoken language is also suitable (if not more) for such kind of shallow processing.

Information Extraction

Information extraction is a process for the automatic extraction of structured or semistructured information from unstructured machine-readable (text) documents.

Information extraction is either rule-based such as in GATE [252] and in the DFKI system SPROUT [253] or based on machine learning methods that mostly involved supervised classifier training over manually annotated corpora.

Information extraction from textual content is situated between information retrieval and text understanding. Unlike information retrieval where the aim is to locate passages of text relevant to a domain-specific topic or a user's query (e.g. news on pole-vault events), information extraction aims to locate inside a text passage domain-specific and pre-specified facts (e.g. facts about the athlete

participating in the pole-vault event, such as his/her name, nationality, performance, as well as facts about the specific event, such as the event name, location). Unlike text understanding, only a small portion of a text is typically relevant to an extraction task.

Information extraction (IE) can be defined as the automatic identification of selected types of entities, relations or events in free text. More specifically, IE is about locating in a text the following different types of information: named entity recognition, semantic tagging and event recognition.

The MUMIS project at Sheffield University has studied how to extract such information across multiple, multimodal sources.

Named Entity Recognition

Named entity recognition (NER) (also known as entity identification (EI) and entity extraction) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. For example, a NER system producing MUC-style output might tag the sentence,

Jim bought 300 shares of Acme Corp. in 2006.

<ENAMEX TYPE="PERSON">Jim</ENAMEX> bought <NUMEX TYPE="QUANTITY">300</NUMEX> shares of <ENAMEX TYPE="ORGANIZATION">Acme Corp.</ENAMEX> in <TIMEX TYPE="DATE">2006</TIMEX>.

NER systems have been created that use linguistic grammar-based techniques as well as statistical models. Hand-crafted grammar-based systems typically obtain better results, but at the cost of months of work by experienced linguists. Statistical NER systems typically require a large amount manually annotated training data.

Since about 1998, there has been a great deal of interest in entity identification in the molecular biology, bioinformatics, and medical natural language processing communities. The most common entity of interest in that domain has been names of genes and gene products.

Although NE extraction from well-formatted text input has been intensively investigated and achieved satisfactory performance, NE extraction from speech remains under-explored. The task of proper names recognition is much more complex than standard keywords recognition task. The number of proper names could be hundred thousand names [254] , [255]. The proper names recognition task is also characterized by very high perplexity factor. One of the main problems in proper names recognition is pronunciation modelling [256], [257], [258]. In the list of proper names could be foreign names. The speaker not always knows how to pronounce correctly foreign name. A lot of examples of not correct pronunciation of foreign spoken names in commentator speech can be seen in the sport recording of large international sport events. Automatic speech recognition system should model all possible pronunciation to capture correctly spoken proper name.

Kubala et al. [259] and Miller et al. [260] applied text-based NE tagger on the first best hypothesis of English broadcast news speech recognition systems, and they noticed that on average 1% WER costs 0.7 points of NE extraction F-score. Palmer et al. [261] extracted NEs directly from word recognition lattices. Zhai et al. [262] applied weighted voting on Nbest hypotheses for Chinese broadcast news speech NE extraction.

Semantic tagging

A separated but related step is the identification of concepts in texts. For this purpose, terms that are indicators of such concepts need to be extracted and semantically classified i.e. mapped to ontology classes that define the concepts. Semantic tagging has been mostly implemented by use of the Princeton Wordnet for English or of wordnets for other languages as defined by the Global Wordnet Association.

Ontologically described information is a basic requirement for more complex processing tasks such as reasoning and discourse analysis. More in particular, there are three main reasons for formalizing

extracted information with respect to an ontology - for related work see e.g. [263], [264], [265], [266], [267]:

- **Architecture**: The SmartWeb system is based on the representation of information with respect to an ontology. Results from different components are represented in a uniform way according to the SWIntO ontology, such that it makes no difference for the central SmartWeb dialog system where the information has actually come from, i.e. from open domain question answering, the knowledge base or from a semantic web service. Complying with the ontology therefore allows for a smooth integration of the information from different processing chains.
- **Information Integration**: Representing information with respect to an ontology and storing it in a knowledge base allows for linking different types of information in a well-founded way, establishing connections between extracted entities and events at the semantic level.
- **Reasoning**: Using a formal ontology allows for applying standard inference engines for reasoning over extracted facts (i.e. entities, events), thus enabling the derivation of further information that is not explicitly contained in the text - in SmartWeb the OntoBroker system is used for inference and reasoning [268].

The formal discussion on associating text documents with Ontologies is presented in Section 4.6. In the following section, the MPEG – 7 Linguistic description scheme is presented as this standard will be used to interoperate with semantic concepts defined using semantic web technologies such as RDFS/OWL, etc.

4.2.2.3.1 MPEG-7: The Linguistic Description Scheme (LDS)

MPEG-7 standard can be used for metadata description. It is an excellent choice for describing audiovisual content because of its comprehensiveness and flexibility. The comprehensiveness results from the fact that the standard has been designed for a broad range of applications and thus employs very general and widely applicable concepts. The standard contains a large set of tools for diverse types of annotations on different semantic levels. The flexibility of MPEG-7, which is provided by a high level of generality, makes it usable for a broad application area without imposing strict constraints on the metadata models of these applications.

The flexibility is very much based on the structuring tools and allows the description to be modular and on different levels of abstraction. MPEG-7 supports fine grained description, and it is possible to attach descriptors to arbitrary segments on any level of detail of the description.

Among the descriptive tools developed within the MPEG-7 framework, one is concerned with the use of natural language for adding metadata to the content description of image and video: the so-called Linguistic Description Scheme (LDS).

MPEG-7 foresees four kinds of textual annotation that can be attached as metadata to some audio-video content. The natural language expression used here is "Spain scores a goal against Sweden. The scoring player is Morientes".

Free Text Annotation: Here only tags are put around the text:

<TextAnnotation>

<FreeTextAnnotation xml:lang="en">

Spain scores a goal against Sweden.

The scoring player is Morientes.

</FreeTextAnnotation>

</TextAnnotation>

Key Word Annotation: Key Words are extracted from text and correspondingly annotated:

<TextAnnotation>

<KeywordAnnotation>

<Keyword>score</Keyword>
<Keyword>Sweden</Keyword>
<Keyword>Spain</Keyword>
<Keyword>Morientes</Keyword>
</KeywordAnnotation>
</TextAnnotation>

Structured Annotation: Question/Answering like semantics is associated to the text:

<TextAnnotation>
<StructuredAnnotation>
<Who><Name>Spain</Name></Who>
<WhatAction><Name>score
goal</Name></WhatAction>
<Where><Name>A Coruña,
Spain</Name></Where>
<When><Name>March 25,
1998<Name></When>
</StructuredAnnotation>
</TextAnnotation>

Dependency Structure: Here the full linguistic apparatus is used for annotating the text:

<TextAnnotation>
<DependencyStructure>
<Sentence>
<Phrase operator="subject">
<Head type="noun">Spain</Head>
</Phrase>
<Head type="verb" base-
Form="score">scored</Head>
<Phrase operator="object">
<Head type="article noun">a
goal</Head>
</Phrase>
<Phrase>
<Head
type="preposition">against</Head>
<Phrase>
<Head>Sweden</Head></Phrase>
</Phrase>

</Sentence>

</DependencyStructure>

</TextAnnotation>

Research Focus in Papyrus

Content analysis will be based on two different approaches: a document driven approach and an ontology driven approach.

The document driven approach aims at identifying the relevant information starting from the collection of documents, without external knowledge. In textual document, POS tagging followed by the evaluation of weighted occurrences of noun, verbs and adjectives, together with identification of chunks, is a first step towards keyword extraction and selection. In spoken document, also prosodic, structural and discourse features, as well as lexical features, are predictors of those audio segments that contain the most relevant information. All of these elements will be taken into account in the first explorative phase, in order to identify the content that is pertinent to Papyrus.

The Information Extraction (IE) task, on both textual documents and transcripts of spoken documents, will provide the instances to populate ontologies and will be based on machine learning techniques, as these have proven to be easier and faster to port to new domains, compared to systems that use hand-crafted patterns and rules. However, the best results are sometimes achieved by hybrid approaches that combine knowledge engineering and machine learning.

In the ontology driven approach, the extracted information is interpreted according to a model of the domain-specific knowledge. We could then characterise IE as an ontology-driven process, since an ontology is a formal description of conceptual knowledge for a specific domain. As noted in [269], IE needs ontologies as part of the understanding process whereas IE can be used, on the other hand, for populating and enriching the ontology. That's why there is recently a trend towards bootstrapping approaches in information extraction, where an IE system starts with seed ontology and, in the course of the process; the extracted information is used to populate the ontology improving in turn the performance of the IE system which exploits the populated ontology.

Research in Papyrus will investigate the contribution of Ontologies to existing IE systems. The aim is to identify a concrete methodology for ontology based information extraction exploiting all levels of ontological knowledge, from the domain entities for named entity recognition, to the use of conceptual hierarchies for pattern generalisation, to the use of properties and non-taxonomic relations for pattern acquisition, and finally to the use of the domain model itself for integrating extracted entities and instances of relations, as well as for discovering implicit information and detecting inconsistencies.

Evaluation will be based on the metrics of Precision, Recall and F-Measure, but other aspects may also be examined such as the closeness of the answer to the correct position in the ontology. With this respect, in [270] below the Augmented Precision and Recall measures are proposed that take into account the ontological distance of the response to the position of the key concepts in the hierarchy. The difference from traditional Precision and Recall is due to the fact that these augmented measures consider weighted semantic distance in addition to a binary notion of correctness.

4.3. Intelligent Relevance Feedback

With the advances in computer technologies and the advent of World Wide Web (WWW), there has been an explosion in the amount and complexity of digital data being generated, stored, transmitted, analyzed and accessed. Much of this information is multimedia in nature, which includes digital images, video, audio, graphics and text data. In order to make use of this vast amount of data, efficient and effective techniques to retrieve multimedia information based on its content needs to be developed. Key word annotation is one of the traditional content retrieval paradigms. In this approach, the images (in this section the word image includes Key-frames[96] extracted from the videos) are first annotated manually by keywords. After which they are retrieved by their corresponding annotations. However, there are three main difficulties with this approach, i.e. the large amount of manual effort required in developing the annotations, the differences in interpretation of multimedia content, and the inconsistency of the keyword assignment among different indexes [271], [272], [273]. As the size of multimedia content repositories increases, keyword annotation approach becomes feasible. To overcome the difficulties of the annotation based approach, an alternative mechanism, content – based image retrieval (CBIR) was proposed in the early 1990's. Despite the extensive research effort, the retrieval techniques used in CBIR systems lag behind the corresponding techniques in today's best text search engines. At the early stage of CBIR, research primarily focused on exploring various feature representations, hoping to find a best representation for each feature. The corresponding system design strategy for early CBIR systems is to first find the best representations for the visual features.

During the retrieval process, the user selects the visual features that he or she is interested in. In the case of multiple features, the user also needs to specify the weights for each of the features, based on the selected features and specified weights, the retrieval system tries to find the similar images to the user's query. From different experiments, it was found out that performance of such systems is not satisfactory due to the following two reasons. One being the gap between high level concepts and low level features, the assumption that the computer centric approach makes is that the high-level concepts to low-level features mapping is easy for the user to do. While, in some cases, the assumption is true, in some other cases, this may not be true and the other being subjectivity of human perception. Different persons, or the same person under different circumstances, may perceive the same visual content differently. This is called human perception subjectivity. The subjectivity exits at various levels. One of the techniques to incorporate user subjectivity is through "Relevance Feedback (RF)". The main idea of RF consist in choosing important features of certain previously retrieved items that have been identified as relevant by the users and emphasis these feature the in a new query formulation Additionally the irrelevant features can be de-emphasized in the future query formulations. This has as effect the alteration of query closer in the direction of relevant items and further away from non-relevant items. The expectation is that more relevant items are retrieved in subsequent search iterations. The RF mechanism provides additional advantages for a retrieval system. The most significant of these are:

- It acts as a conceptual screen between the user and the query formulation mechanism, allowing the user to formulate powerful queries without intimate knowledge of the search process or of the archive structure.
- It structures the search process by breaking the search operation into sequences of iterative steps designed to gradually approach the targeted relevant documents.

It provide a controlled environment for query formulation and subsequent adaptation by allowing the user to emphasis relevant items and theirs features as required by the particular information needs of the users.

RF introduces human visual perception into the retrieval process gradually, is an efficient improvement for narrowing down the gap between low-level visual feature representation of an image and its

[96] A key-frame is a most representative image of a shot extracted from the video.

semantic meaning in content-based image retrieval (CBIR). Without detailed knowledge of the video archive structure, and of the retrieval environment, most users find it difficult to formulate well-designed queries. Since the query formulation process is not transparent to the retrieval system users the initial query is likely to be far from an optimal formulation. Consequently the initial retrieval operation can be considered as being a trial run only designed to retrieve a few useful items from a given collection. The items retrieved in the in the initial run can then be examined for relevance and the query formulation adapted accordingly in the hope of retrieving additional useful items during subsequent search operations.

In this section, we will present the state of the art techniques for relevance feedback based on visual features extracted from either images or sequence of images (video).

Rui et al., in [275] proposed a relevance feedback approach to CBIR to overcome the problems faced by computer centric approach. In relevance feedback a human and a computer interact to refine high-level query representation based on low-level features. Visual feature extraction applied to content-based image retrieval has been thoroughly studied for last decade. Most work concentrates on low level visual features such as colour, shape, texture, and adopts a feature-based image retrieval approach. The application of these systems to real world problems is however limited due to the ignorance of content varieties and the lack of semantic meanings in extracted features. Specific low-level image features may provide a solution to image retrieval in some applications (e.g. with respect to a pre-selected image database), but may have a problem in handling other applications because there isn't neither a universal feature applicable for all images. Hence, to improve the image retrieval performance, it becomes vital to integrate image classification tools with the feature-based retrieval techniques. The image retrieval engine usually consists of a human – user interface, an image analysis unit and a matching mechanism [276]. The image analysis part in existing CBIR systems is fairly simple, i.e., to extract features specified by users. In an effort to overcome semantic gap, CBIR systems include humans in the retrieval process. In the following section, a generic overview of the Relevance Feedback system is presented followed by different learning techniques. The section concludes with an overview of research scope of Relevance Feedback within Papyrus.

Overview of Relevance Feedback Systems

Query Vector

The relevance feedback was originally designed for text retrieval where the query model consists of a weighted selection of search terms [277], [278] and [279]. A query vector can be then written as:

$$Q_0 = (q_1, q_2, ...q_i) \tag{4.3.1}$$

Where q_i represents the weight of term i in query Q_0. The weights are in the range between 0 and 1; with 0 representing a term absent from the query vector and 1 represents a full weighted term. A term could be a word chosen from a term dictionary or even a full phrase in the natural language of the user. From an initial query vector the relevance feedback derives an updated vector:

$$Q_0' = (q_1', q_2', ...q_i') \tag{4.3.2}$$

Where q_i' represents the altered term weight assignments for the i index terms. New terms are introduced in the query by assigning them with a positive weight older terms are removed by reducing their weigh to 0. In this approach the feedback process can be visualized as a shift in the query vector from one area to another into the $T-$ dimensional space defined by the T index terms.

Basic feedback

Both the information items D stored in the collection and the requests for information Q can be represented as $T-$ dimensional vectors of the form:

$$D_0 = (d_1, d_2, ..., d_i) \tag{4.3.3}$$

$$Q_0 = (q_1, q_2, ..., q_i) \tag{4.3.4}$$

Where d_i and q_i represent the weight of term i in D and Q, respectively. The query-document similarity measure can then be computed as the inner product between corresponding vectors:

$$Sim(D,Q) = \sum_{i}^{T} d_i \cdot q_i \tag{4.3.5}$$

The optimal query is a set of query documents for which best retrieval results are obtained and is denoted as $Q_{optimal}$ for the above given similarity is of the form [280],

$$Q_{optimal} = \frac{1}{n} \sum_{relevant} \frac{D_i}{|D_i|} - \frac{1}{N-n} \sum_{non-relevant} \frac{D_i}{|D_i|} \tag{4.3.6}$$

Where D_i represents the document vectors, $|D_i|$ is the corresponding Euclidean vector length, N is the size of the collection and n the number of relevant documents in the collection. However the above optimal query cannot be used in practice as an initial query formulation because the set of n relevant documents is not known in advance. The optimal query is employed in generating a feedback query once relevance assessments are available for some of the items previously retrieved in the initial search iteration. In this case the updated query following the retrieval of n_1 relevant and n_2 non-relevant items can then be formulated as:

$$Q_1 = Q_0 + \frac{1}{n} \sum_{known\ relevant} \frac{D_i}{|D_i|} - \frac{1}{n_2} \sum_{known\ non-relevant} \frac{D_i}{|D_i|} \tag{4.3.7}$$

Where Q_0 and Q_1 represent the initial and first iteration queries respectively. In the general formulation the expression can be written as:

$$Q_{i+1} = \alpha Q_i + \beta \sum_{\substack{known \\ relevant}} \frac{D_i}{|D_i|} - \gamma \sum_{\substack{known \\ non-relevant}} \frac{D_i}{|D_i|} \tag{4.3.8}$$

Where the normalized weights α, β and γ are between 0 and 1. This vector alteration approach is conceptually simple; the modified term weights being directly obtained from the weights of the corresponding terms in relevant and non-relevant documents. When the weights accurately reflect the real values of the terms standard vector modification process provides a powerful query construction method. Relevance feedback in image retrieval is in general a supervised image classification problem representing effort towards bridging the semantic gap between automatically extracted visual features and predefined semantic categories as relevant and irrelevant.

Relevance Feedback Learning Techniques

Relevance Feedback (RF) techniques highlight the importance of learning methods in CBIR. Learning has indeed been the dominating factor to narrow the semantic gap arising from the low-level feature representation in the last few years. The idea of incorporating relevance feedback first emerged in text retrieval systems [281], and has been studied since. In comparison to pure text IR systems, relevance feedback is even more valuable in the image domain: a user can tell instantaneously whether an image is relevant with respect to their current context (information need, awareness of information need, etc.), while it takes substantially more time to read through a text document to estimate its relevance.

Relevance feedback is regarded as an invaluable tool to improve CBIR systems, for several reasons. Apart from providing a way to embrace the individuality of users, they are indispensable to overcome the *semantic gap* between low-level image features and high-level semantic concepts. The user's judgement of relevance is naturally based on their current context, their preferences, and also their way of judging the semantic content of the images. By prompting the user for relevance feedback, the initial estimation could be improved to steer the results in the direction the user has in mind. Rather than trying to find better techniques and more enhanced image features in order to improve the performance of what has been referred to as *"computer-centric"* systems [275], it is more satisfactory

to the user to exploit human computer interaction to refine high level queries to representations based on low level features. This way, the subjectivity of human perception and the user's current context are automatically taken into account, as well. Consequently, it does not come as a surprise that there exist various techniques of how to make use of relevance feedback in CBIR.A comprehensive study of existing relevance feedback techniques in image retrieval can be found in [282]has represented a comprehensive study on existing feedback technique in image retrieval.

Relevance feedback is engaged with finding optimised ways of updating the parameters of the retrieval algorithm. Traditionally, this has been achieved through query refinement approaches [283], [284], [285]. These approaches underlie a geometric interpretation of the feature and query space. In most CBIR systems, the images are represented by their feature vectors in the vector space model [286]. So, query refinement approaches strive to find the "ideal" query point that minimises the distance to the positive examples provided by the user. Alternatively, relevance feedback has also been formulated in Bayesian frameworks as belief propagation [287], [288], or as a classification task [289], [290].

Neural Network based Relevance Feedback

One of the techniques for integrating learning approaches in relevance feedback is Neural Network training. The neural network based relevance feedback is based on Self Organizing Maps (SOM), or Self Organising Feature Maps (SOFM), thus interchangeably used. In [291] the authors present a relevance feedback model based on an associative neural network in which meaningful concepts to the user are accumulated at retrieval time by an interactive process. The network was regarded as a kind of personal thesaurus to the users. A rule based superstructure is then defined to expand the query evaluation with the meaningful terms identified in the network. The search terms are expanded by taking into account their associations with the meaningful terms in the network. The authors apply this approach to Information Retrieval, which in general performed through an iterative and cooperative process of trial and error between the user and the system. In this approach the authors generate the thesaurus of concepts on the basis of the relevant documents selected by the user from among those retrieved by the original query. The number of documents which can be selected in this phase must be small enough to guarantee acceptable relevance feedback performances; for this reason, a fixed maximum number of relevant documents should be set for each application. The user must specify at least one document among those retrieved that he/she judges as fully relevant to his/her needs.

In [292] authors propose to address the issue of image retrieval which corresponds to human perception. The authors propose to control the order vector used in synergetic neural nets (SNN) and use it as the basis of a similarity function for shape based retrieval. Based on the properties an efficient affine invariant similarity measure has been developed for trademark images. Furthermore a self-attentive retrieval and relevance feedback mechanism for similarity measure refinement is presented.

In [293], a neural network scheme is presented for adaptive video indexing and retrieval. In this approach, a limited but characteristic amount of frames are extracted from each video scene, able for providing an efficient representation of the video content. For this reason, a cross correlation criterion is minimized using a genetic algorithm. Low level features are extracted to indicate the frame characteristics, such as colour and motion segments. After the key frame extraction, the video queries are implemented directly on this small number of frames. To reduce the limitation of low-level features, the human consideration is incorporated to assign a degree of appropriateness for each retrieved image of the system and then restart the searching. A feedforward neural network structure is proposed as a parametric distance for the retrieval, mainly due to the highly non-linear capabilities. An adaptation mechanism is also proposed for updating the network weights, each time a new image selection is performed by the user. The algorithm results in a convex minimization and thus a minimum always exists.

The PicSOM system [294] is content based image retrieval system, which uses SOM for indexing images with their low-level features. SOM's represent unsupervised topologically ordered neural networks, which project a high-dimensional input space (n-dimensional feature vectors) into a low-

dimensional lattice. The latter, usually being a two-dimensional grid with n-dimensional neighbours connected in appropriately weighted nodes.

In [295], a fuzzy RF approach is introduced, in which the user provides a fuzzy judgement about the relevance of an image, unlike in binary relevance systems with a hard decision on relevance. A hierarchical tree with multiple levels of informational provided to the user is defined in the first step. A continuous fuzzy membership function is used to model user's fuzzy feedback by weighting different images labelled as fuzzy with different weights to simulate user's perception. For learning users preferences and visual content interpretation a radial basis function (RBF) neural network is used. RBF neural network in combination with the fuzzy approach is denoted as fuzzy radial basis function (FBRF) network.

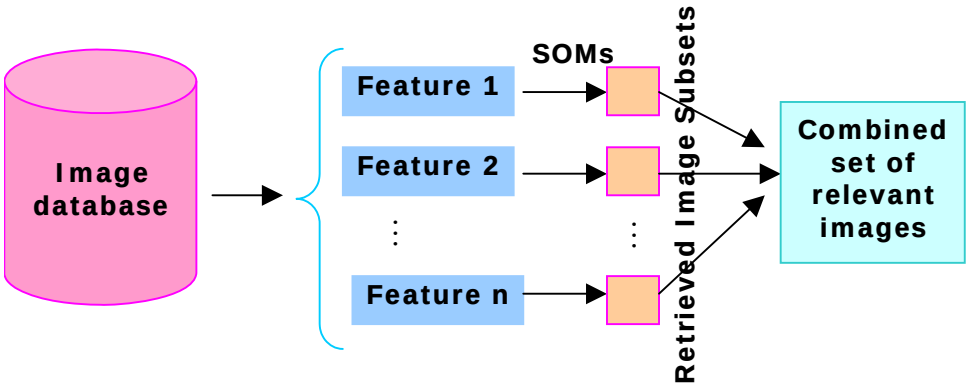

Figure 4-2: Architecture of the PicSOM system

It has one input layer, one output layer and three internal modular sub-networks constituting of a Gaussian kernel layer and relevance contribution layer. These modules are associated with "positive", "negative" or "fuzzy" user's feedback. An input to the first layer is a feature vector of an image and the output layer combines responses from the modules as a liner combination of three sub-networks.

When creating FRFB we have these sets of training feature vectors, relevant $V_r = \{v_1, ..., v_R\}$, irrelevant $V_{ir} = \{v_1, ..., v_{IR}\}$ and fuzzy $V_f = \{v_1, ..., v_F\}$ and a set of all training samples $V = \{v_1, ..., v_M\}$ of dimension M. The elements v_i represent P dimensional feature vectors. The Gaussian shaped RBF modules are defined with a function:

$$f_i(x, v_i) = \exp\left(-\frac{(x - v_i) \wedge (x - v_i)}{2\sigma_i^2}\right), \; i = 1, ..., M \tag{4.3.9}$$

Where x is the input feature vector, v_i, σ_i are mean and variance of an appropriate $i-$th module, and $\wedge$ is a diagonal matrix of dimension P with elements representing importance of feature components.

As for the weights between the Gaussian kernel and relevance contribution layer they are for all positive examples equal and positive, whereas for negative examples equal and negative. As for fuzzy samples weights are determent using a fuzzy membership function based on a closeness of a sample to the centre c_γ of a relevant cluster. In [296] Cauchy function is used as a membership function, with compactness of a class denoted as τ and shape (smoothness) of the function as γ:

$$w_f^i(v_i) = \frac{1}{1 + \left(\frac{\| v_i - c_\gamma \|^\gamma}{\tau}\right)} \tag{4.3.10}$$

Therefore, the overall similarity measure it is a linear combination of outputs from each module:

$$F(x) = \sum_{v_i \in V_r} w_\gamma^i f\gamma(x, vi) + \sum_{v_i \in V_{ir}} w_{ir}^i fir(x, vi) + \sum_{v_i \in V_f} w_f^i f_f(x, v_i) \tag{4.3.11}$$

Accordingly if feature vector x is close to the positive class centre v_i the output is large and the images are similar, or if it's close to negative the output is a large negative value thus a small value and the image is less similar.

Bayesian Framework based Relevance Feedback

In [297], a Bayesian decision theory is introduced to estimate the boundary between relevant and non-relevant images in RF mechanisms. The Bayesian approach computes the new query point based on relevance feedback from the user scenarios where images are indexed by a global feature vector, the similarity function is defined through a metric measure, and images are retrieved by the k-nn algorithm. The basic idea is to use local estimation of the decision boundary between the "relevant" and "non-relevant" images in the neighbourhood of the original query. The new query is then placed at a suitable distance from such boundary, on the side of the region containing relevant images.

In [298], a new relevance feedback approach based on Bayesian classifier is proposed. This approach treats positive and negative feedback examples with different strategies. Not only can the retrieval performance be improved for the current user, but the improvements can also help subsequent users. The authors also apply the Principle Component Analysis (PCA) techniques, the feature subspace is extracted and updated during the feedback process, so as to reduce the dimensionality of feature spaces, reduce noise contained in the original feature representation and hence to define a proper subspace for the type of feature as implied in the feedback. These are performed according to positive feedbacks and hence consistently with the subjective image content. To incorporate positive feedback in refining image retrieval, the authors assume that all of the positive examples in feedback iteration belong to the same semantic class whose features follow Gaussian distribution. Features of all positive examples are used to calculate and update the parameters of its corresponding semantic Gaussian class and we use a Bayesian classifier to re-rank the images in the database. To incorporate negative feedback examples, the authors apply a penalty function in calculating the final ranking of an image to the query image.

The parameters for a semantic Gaussian class can be estimated using the feature vectors of all the positive examples. Hence, the image retrieval becomes a process of estimating the probability of belonging to a semantic class and the query refinement by relevance feedback becomes a process of updating the Gaussian distributing parameters. The log posterior probability that the feature vector x belongs to the semantic class c_i implied in the positive examples is estimated using the following bayesian formulation.

$$g_i(x) = \ln(P(c_i \mid x) \alpha \ln p(x \mid c_i) + \ln P(c_i)$$
$$g_i(x) = -\frac{1}{2}\frac{(x-\bar{x}_i)T)}{\sum_i (x-\bar{x}_i)} + \ln P(c_i) + const_i \qquad (4.3.12)$$

Where $const_i$ are fixed in one feedback iteration. Assuming that different feature types are independent of each other, the log posterior probability is calculated as the sum of the individual $g_i(x)'s$. Experiments show that the use of this posterior probability as the ranking metric improves the performance of relevance feedback in content-based image retrieval. There are three parameters in the Bayesian estimator (5). They need to be updated when more positive examples are provided by the user through relevance feedback. Actually such a process could be considered as the combination of two Gaussian classes. So it is easy to get the updating process. Denote the current set of positive examples by U , in which each example is denoted by $\underline{u}$. In other words, there are totally $|U|$ positive examples in the current feedback iteration. The updating of the Gaussian parameters is performed as follows:

$$n' = n+ |U|$$

$$\sigma_i^{'2} = n\sigma_i^2 + \frac{n|U|\bar{x}_i^2 - 2n\bar{x}i\sum_{u\in U} u}{n+|U|} + \sum_{u\in U} u^2 - \frac{\left(\sum_{u\in U} u\right)^2}{n+|U|} \quad (4.3.13)$$

$$\bar{x}_i^{'} = \frac{n\bar{x}i + \sum_{u\in U} u}{n+|U|}$$

Where n and n' are the total number of positive examples accumulated before and after the current feedback iteration, respectively; $\sum_{u\in U} u$ represents the mean of the positive examples in the current iteration.

In the implementation carried out, the probability of a semantic class $P(c_i)$ is assumed equal for all semantic classes and remains constant in the relevance feedback process. The following describes how the positive feedback is performed.

- Initialization of System:

Feature Normalization: This allows equal emphasis on all the feature components. For $\vec{x}_i^{'} = [x'_{il},..., x'_{ini}]$ where $x'_{im} = (x'' - \bar{x}(x''_{im})/3\sigma(x_{im}))$ and $x''_{i_m} = (x_{i_m} - \min(x_{i_m})/\max(x_{i_m}) - \min(x_{i_m}))$. If x_{i_m} satisfies the Gaussian distribution, it is easy to prove that the probability of x'_{i_m} being in the range of [-1,1] is 99%.

Initialization: Initialize $\sigma_i = I(identity\ matrix)$ and $\bar{x}' = \vec{x}_i, n = 1$.

- Retrieval and Feedback:

Update the retrieval parameters σ_i, x_i and n according to the equations below using the information provided by the current set of positive example U.

Distance Calculation: For each image K, $K \in D$ its distance d_i is calculated using equation in the retrieval after the feedback $d_i = -g(K)$.

- Sorting by distances and provide the new ranking list to the user.

Most methods apply the same methodology to negative and positive examples based on the assumption that the negative examples have the same feature distribution as the positive ones; or otherwise ignore the negative examples completely in the feedback process. In the approach the authors proposed in [299], the negative examples are often isolated and independent. Thus they need to be treated differently from the positive examples. The authors punish those images in the database that are very near to the negative samples and do not let the negative samples influence the other images. Under this strategy, the penalized images near the negative examples by increasing their distance to the query. Such a penalty function seems like a 'dibbling' process in feature space. The penalty function is approximated by a Gaussian function for each negative example. Denote the current set of negative examples by V for each K. The distance punish function is defined as

$$pun(d_i) = \sum_{v\in V} g_v(d_v) \quad (4.3.14)$$

Where $g_v(d_v)$ the Gaussian function is whose parameters is determined experimentally; and d_v is the distance between image K and a negative example v. d_v can be calculated using the Euclidean distance between the feature vectors of image K and the negative example. $\sum_{v\in V}$ sums up penalty contributions from all negative examples to the image K. That is, if an image in the database is close to all negative examples, the penalty is high; in contrast if the image is far away from all negative examples, the penalty function will be decreased to zero, according to the Gaussian distribution. So

the distance after negative feedback is defined as $d_i' = d_i + pun(d_i)$. That is, if an image in the database is close to negative examples, its distance to the query is increased by $Pun(d_i)$.

In [300], the authors presented a generalized Bayesian framework for relevance feedback in content based image retrieval. The proposed feedback technique is based on Bayesian learning method and incorporates a time-varying user model into the formulation. The authors define the user model with two terms: a target query and a user conception. The target query is aimed to learn the common features from relevant images so as to specify the user's ideal query. The user conception is aimed to learn a parameter set to determine the time-varying matching criterion. Therefore, at each feedback step, the learning process updates not only target the distribution but also the target query and the matching criterion. Also, the relevance feedback model presented works on the region based image representations. The matching criterions are formulated using a weighting scheme and a region clustering technique to determine the region correspondence between the relevant images.

SVM based Relevance Feedback

Support Vector Machines (SVM) based relevance feedback falls under the category of Discriminative Classification Models which do not try to describe classes but the boundaries separating these classes. This category also includes Fisher's Discriminative Analysis (FDA). Relevance Feedback based SVM provides a supervised learning method, describing hyperplanes in feature space that separate classes [301], [302]. In [303] authors use a combination of weighted retrieval system with Mahalanobis distance as a similarity measure and SVM for estimating the weight of relevant images in the covariance matrix. This approach is a combination of already exploited techniques and new statistical learning algorithm SVM. The overall similarity for a particular image in the database is obtained by linearly combining similarity measures for each feature, as in other approaches previously discussed:

$$S_j = \sum_i W_i S_j(f_i),\ j = 1,\ldots, N \tag{4.3.15}$$

Where N is the number of images in the database, f_i individual feature of an image and S_j is the Mahalonobis distance (2.7), used as a similarity measure. Weights for the lower level features in Section 4.1 are updated as follows:

$$di = \frac{\sum_{k=1}^{NR} V(k) S_k(f_i)}{\sum_{k=1}^{NR} V(k)},\ W_i = \frac{1}{d_i} \tag{4.3.16}$$

Where $V(k)$ denotes the weight for the $k-$th relevant image, which will be determined by the use of SVMs, NR represents the overall number of positive feedback examples. A smaller normalized distance d_i gives a higher weight to a feature.

For determining the weight for k -th relevant image $V(k)$ users feedback SVMs are exploited. The aim is to separate and classify positive and negative examples. For this purpose the user must give negative and positive example feedback, and the new weights for relevant examples are automatically obtained from SVM learning.

If a user provides a set of training samples either positive +1 or negative –1:

$$\{(\vec{x}_i, y_i)\}_{i=1}^{N},\ y_i = +1/-1 \tag{4.3.17}$$

Where $\vec{x}_i$ is a feature vector for the i -th image and y_i is a label. SVM searches for an optimal hyperplane to separate positive and negative examples:

$$w^T \vec{x} + b = 0, \tag{4.3.18}$$

Here w denotes the weight and b the bias. SVM tries to find optimal w_0 and b_0 to maximize the distance between the feature vectors belonging to different classes but being closest to the separation hyperplane. Thus the distance of a point from the plane is given as:

$$d(w_0, b_0, \vec{x}) = \frac{| w_0{}^T \vec{x} + b_0 |}{\| w_0 \|} \quad (4.3.19)$$

Different weights are assigned based on the distance of positive examples from the hyperplane, the larger the distance the more distinguishable the examples are from the negative ones and the large the weights. In case that the two classes are not linearly separable implementing inner product kernel functions maps the input feature space into a higher dimensional space where the boundary can be easily determined.

A new approach proposed in [304], [305] is a region-based method for extracting local region features. Automatic extracting of semantically meaningful image objects is still not fully possible even in state-of-the-art segmentation methods. Some RF approaches partition an object into several regions and ask the user to determine relevant one, in this way they place an additional burden on the user. Whereas in this approach the authors combine regions and perform image-to-image similarity matching by using EMD (described in section 2), which allows different dimensions of feature vectors. In SVM-based classification both positive and negative labelled images are used as training data to learn the classifier how to separate the unknown part of the database, the test set, into two or more classes. Kernels in SVM are based on inner product in the input space, and in [306] a new kernel is introduced to better accommodate region-based approach. This new kernel is a generalization of the Gaussian with the Euclidian norm replaced by EMD [307]:

$$K_{Gaussian}(x, y) = \exp(-d(x, y) / 2\sigma^2) \quad (4.3.20)$$

Where d is the distance measure and in the general case this is Euclidian norm, in this specific case it is EMD. Signature for EMD for each image is a vector of pairs (feature of a region, weight of a region) for all regions. Hence, the length of image representations is variable due to different numbers of segmented regions. EMD incorporates features of all the regions allowing many-to-many relationships and thus robustness to inaccurate segmentations [308].

Other techniques

In [315] the authors focus on effective feature space dimension reduction according to user's feedback but also to improve the image description during the retrieval process by introducing new significant features. The approach relies on an effective combination of query and feature weights refinement, which are simple and computationally effective, with a new concept of semantically – based feature based space modification, thus achieving a feature-adaptive relevance feedback (FA-RF), able to automatically tune the feature space to the user's perception of image similarity, while maintaining a compact image description. A transformation of the original features is derived from the user's input that improves the retrieval performance.

In [316] the authors present an approach for active concept learning in image databases. In general, the methods presented earlier, discuss relevance feedback for each individual user session. The algorithms do not learn concepts in the feature space systematically. Furthermore, once the user is done with a query and starts a new query, the meta knowledge gained by the system with previous queries is lost. Meta knowledge is the experience of each query image with various users. This experience consists of the classification of each image into various classes (clusters), relevance's (weights), of feature vectors and the number of times this image is selected as a query and marked as positive and negative. Since, real image databases experience retrievals from many users, it is possible to exploit previous retrieval experiences (meta knowledge) to learn and refine visual concepts. In [317] authors combine traditional relevance feedback methods with the technique of virtual feature which is derived from long-term retrieval experience. The image dissimilarity measure can be adapted dynamically depending on the estimate of the relevance probability derived from the virtual feature. In [318] the authors combine users' annotation in parallel with the content-based similarity which is called compound query technique. In [319] authors integrate feature relevance

learning with fuzzy clustering which partitions the image dataset for efficient indexing with the help of meta knowledge. In image retrieval research, few statistical models have been developed for strict theoretical analysis of concept learning. In [320], [321], authors analysed the probabilistic image retrieval model based on mixture densities for the quality of the solution and computational complexity. However, the exploitation of meta knowledge is not considered for the model. In [322], authors organised images (with associated words) by a hierarchical model for browsing and searching. In their work, some images are used to train the clustering directly; however, this training stage is unreliable since the training data set may not represent the image distribution of the entire database, especially when some images are added or removed during the database lifetime. The task of concept learning is to explore the characteristics of the features that can represent a concept as perceived by various users. Specifically, for guassian mixture model assumption, concept learning estimates the parameters for the Gaussian corresponding to a concept. One of the major approches to estimate (finite) mixture model is to use expectation maximisation (EM) algorithm to estimate mixture model parameters [323]. In [316], authors propose a new active learning approach for mixture model fitting, which includes a model selection method and a user directed semi-supervised EM algorithm. The retrieval experiences derived from previous users' feedback are used to achieve concept learning, which may help to improve future retrieval performance.

In [324], authors propose a nonlinear approach to simulate human perception for relevance feedback. This allows for effectively bridging the gap between the low-level features used in retrieval and the high-level semantics in human perception. The framework uses a specialised radial-basis function (RBF) network [325], [326] for learning the users' notion of similarity between images. In each interactive retrieval session, the user is asked to separate from a set of retrieved images, those which are more similar to the query image from those which are less similar. The feature vectors extracted from these classified images are then used as training examples to determine the centers and widths of the different RBF's in the network. This concept is adaptively redefined in accordance with different users' preferences and different types of images, instead of relying on any preconceived notion of similarity through the enforcement of a fixed metric. Compared to the conventional quadratic measure and the limited adaptivity allowed by its weighted form, the current approach offers an expanded set of adjustable parameters, in the form of the RBF centers and widths. Hence, this allows a more accurate modelling of the notion of similarity from the user's view point.

The approaches presented above perform retrieval based primarily on global features. However, it is not unusual that users accessing a CBIR system look for objects. Thus, the aforementioned systems are likely to fail, since a single signature computed for the entire image cannot sufficiently capture the important properties of individual objects. Although relevance feedback has shown its great potential in improving the retrieval performance in CBIR systems that use global feature representations, it has seldom been introduced to the region-based retrieval systems. In [328], authors presented a pioneering work in this area by proposing the FourEyes system. The system contains three stages: grouping generation, grouping weighting and grouping collection. In the generation stage, it produces many plausible groupings including both with-in image groupings and across-image groupings induced by different models. The collection stage is guided by a rich example-based interaction with the user. The weighting stage adapts the collection stage's search space across users, so that in later interactions, good groupings are found given few examples from the users. In spite of its many good characteristics FourEyes has two disadvantages. One is the use of region-to-region similarity measure, and the other is the re-clustering of all the features when a new image is added. Thus is not very scalable. Other efforts in this direction include IDQS system developed by Wool et al in [329]. In IDQS, a query is initiated by the selection of a region of interest from a key image. After that, user's feedback is given in the form of acceptance or rejection of the retrieved images. Then the Learning Vector Quantisation (LVQ) algorithm is employed to cluster the selected regions in feedback examples. Images with regions close to the positive cluster centroids are then returned and reclassified by the user. This iterative refinement continues until the user is satisfied with the results. Hence, in [327] the authors present a region based image retrieval (RBIR) methods in an attempt to overcome the drawbacks of global features by representing images at object – level, which are closer to the perception of human visual system.

Also, in [330], authors present an object based approach for relevance feedback. Bearing in mind that object segmentation is arguably as hard as the semantic gap problem authors present a block based structure to label single objects in images. As a motivation, authors argue that, labelling complete images as relevant to a given keyword introduces a lot of noise due to the variety of non-relevant objects in complex scenes. The same aspect was studied in a few other approaches that use interest points or models composed of local characteristics of image parts [331], [332], [333]. However, most of these methods are parametric and assume the input data can be faithfully modelled by some probability distribution. Though the advantage of using interest points in an object recognition scenario is apparent, there is guarantee that in a retrieval scenario for natural images the points and features describing the local regions around the points will be representative enough. Natural image databases usually do not contain different views of the same object but a variety of pictures of the same conceptual object. The feature space considered incorporates both the low-level similarity in the multifeature space and spatial correlation among neighbouring blocks. The authors also define a convolution kernel to handle multifeature space. Thus, improving the performance of the retrieval system.

Research Focus in Papyrus

In Papyrus, the research will focus on investigating challenges of contempt context. It is defined as modelling the context under which negative samples (based on metadata model, annotations and low-level features) are selected by the user will enable retrieval of more positive concepts as compared to analysing the positive sets. Contempt here means the dissatisfaction expressed by the user in selecting negative sets from the results provided by the relevance feedback systems. Other research objectives investigated in this activity of Papyrus includes techniques to reinforce automatic annotation, relevance feedback based on semantic user interaction, develop classification models based on machine learning techniques.

4.4. Knowledge Extraction

Early attempts to address "Semantic gap" concentrated on visual similarity assessment via the definition of appropriate quantitative image descriptions, which could be automatically extracted and matched with suitable metrics in the corresponding feature space [334]. Moving from low-level perceptual features to high-level semantic descriptions that match human cognition is the final frontier in computer vision, and consequently to any multimedia application targeting efficient and effective access and manipulation of the available content. As discussed earlier, the early efforts targeting the paved way for content-based (analysis and) retrieval approaches, where focus is on extracting the most representative numerical descriptions and defining similarity metrics that emulate the human notion of similarity. Whilst low-level descriptors, metrics and segmentation tools are fundamental building blocks of any image manipulation technique, they evidently fail to fully capture by themselves the semantics of the visual medium; achieving the latter is a prerequisite for reaching the desired level of efficiency in image manipulation. The limitations of such numerical-based methodologies however, led to the investigation of ways to enhance their performance. The currently developed systems still could not meet realistic user needs, although some have proven particularly effective within certain application context. As a result research focus shifted to the exploitation of implicit and/or prior knowledge that could guide the process of analysis and semantics extraction. In other words, research efforts have concentrated on the semantic analysis of images, combining the aforementioned techniques with a priori domain specific knowledge, so as to result in a high-level representation of images [335], [336]. Domain specific knowledge is utilized for guiding low-level feature extraction, higher-level descriptor derivation, and symbolic inference [337], [338]. Numerous approaches have been proposed building on this principle, exploiting varying methods for modelling this knowledge, varying representations and consequent handling techniques.

Depending on the adopted knowledge acquisition and representation process, two types of approaches can be identified in the relevant literature: implicit, realized by machine learning methods, and explicit, realized by model-based approaches. The usage of machine learning techniques has proven to be a robust methodology for discovering complex relationships and interdependencies between numerical image data and the perceptually higher-level concepts. Moreover, these elegantly handle problems of high dimensionality. Among the most commonly adopted machine learning techniques are Neural Networks (NNs), Hidden Markov Models (HMMs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Genetic Algorithms (GAs) [339], [340]. On the other hand, model-based image analysis approaches make use of prior knowledge in the form of explicitly defined facts, models and rules, i.e., they provide a coherent semantic domain model to support "visual" inference in the specified context [341], [342]. These facts, models and rules may connect semantic and symbolic concepts with other concepts, or with low-level visual features. Regardless of the adopted approach to knowledge representation, the inclusion of spatial information in the knowledge exploited during the analysis process makes necessary the definition and extraction of spatial relations from the visual medium. The relevant literature considers two categories of approaches for the latter task: angle-based and projection-based approaches. Angle-based approaches include [343], where a pair of fuzzy k-NN classifiers are trained to differentiate between the Above-Below and Left-Right relations, and the work of (Millet, 2005), where an individual fuzzy membership function is defined for every relation and applied directly to the estimated angle-histogram. Projection-based approaches include [344], where qualitative directional relations in terms of the centre and the sides of the corresponding objects' MBRs were defined, and [345], where the use of a representative polygon was introduced.

Furthermore, in the real world objects exist in a context. Representing context is a research issue of great importance [346], affecting the quality of the produced results, especially in the field of multimedia analysis in general and knowledge-assisted image analysis in particular. The latter can be defined as a tightly coupled and constant interaction between low level image analysis algorithms and higher level knowledge representation [347]; an area where the role of context is crucial. In recent years, a number of different context aspects related to image analysis have been studied, and a number of different approaches to model context representation have been proposed [348].

Intense past research in the domain of knowledge representation and reasoning with knowledge has, over the last decade, gained new interest in the context of the Semantic Web [334]. New languages such as RDFS (Resource Description Framework Schema) and OWL (Web Ontology Language) have been defined by the World Wide Web consortium (W3C) in order to render meaning to information on the Web and allow for better methods of search and retrieval. As a next step, inference rules and logic are to be used by intelligent applications to derive enriched information from the existing one. Ontologies, which define a set of meanings for a specific domain of information, play an important role in the implementation of the Semantic Web.

Content-based analysis of multimedia requires methods that automatically segment images, video sequences and key frames into areas corresponding to salient semantic objects (e.g. cars, road, people, field, etc), track these objects over time, and provide a flexible infrastructure for further analysis of their relative motion and interactions, as well as object recognition, metadata generation, indexing and retrieval. This problem can be viewed as relating symbolic terms (concepts of the related domain ontology) to visual information by utilizing syntactic and semantic structure in a way similar to approaches in speech and language processing. Moving from low-level perceptual features to high-level semantic descriptions that match human cognition is the final frontier in computer vision, and consequently to any multimedia application targeting efficient and effective access and manipulation of the available content. Limitations of such numerical-based methodologies however, led to the investigation of ways to enhance their performance. Research focus shifted to the exploitation of implicit and/or prior knowledge that could guide the process of analysis and semantic extraction. Numerous approaches have been proposed building on this principle, exploiting varying methods for modelling this knowledge, varying representations and consequent handling techniques. For instance, MPEG-7 compliant low-level multimedia features (e.g. MPEG-7 visual descriptors) can be assigned to semantic concepts thus forming a-priori knowledge base. Processing is then performed by relating high-level symbolic representations to extracted features in the signal (image and temporal feature) domain, thus identifying objects and their relations in the multimedia content. Basing such a representation on ontology, one can capture both concrete and abstract relationships between salient visual properties. The annotation process starts with the segmentation and the extraction of corresponding descriptors (audiovisual features and spatiotemporal relations), only this time the detection and thus annotation is realized as a matching process against the explicitly provided knowledge. Additionally, the use of prior explicit knowledge means that the available knowledge can be used as well to drive the initial steps by determining for example which descriptors that should be extracted or by guiding the segmentation towards more meaningful results.

Machine learning techniques have been applied in many domains, e.g., to learn medical diagnoses, to predict weather, in speech recognition, chemistry, geology. The obvious kind of its applications is associated with knowledge acquisition for Expert Systems, Knowledge Discovery from Databases, and Data Mining. A number of methods have been developed for this type of tasks. The common background of these methods is *similarity based learning*. The basic assumption here is that examples belonging to same class (or segment) can be described using similar characteristics (thus creating clusters in so called feature space). The methods differ in the way in the way how the knowledge is represented (trees, rules, etalons and probabilities), what type of task they can solve (classification, prediction and segmentation), or how complex clusters (classes, segments) they can express (e.g. if the clusters are linearly separable).

Support Vector Machines (SVMs) belong to the general category of machine learning techniques, but due to their good performance in classification problems, are discussed in separately in more detail. Support Vector Machines (SVM) are based on two ideas: (1) we can transform a problem that is not linearly separable in low dimensional feature space into a problem that is linearly separable in high dimensional space, (2) when building a classifier for linearly separable classes, we can consider only those examples, that are closest to the decision boundary. SVM introduce a method that allows us (using so called kernel functions) to use the afore-mentioned ideas without explicitly knowing the transformation from low dimensional into high dimensional space. Other related techniques include neural networks, knowledge discovery in databases and data mining, multimedia mining, relevance feedback, incremental learning as well as hybrid approaches. In addition to the above approaches of knowledge-assisted analysis, multimedia reasoning, based either on implicit knowledge and statistical

methods, or on explicit knowledge and matching processes, can play an important role towards automatic annotation and understanding of multimedia content. By multimedia reasoning in this context, we refer to the automatic derivation of high-level semantic annotations from low-level multimedia data (raw and/or pre-processed to acquire audiovisual or conceptual descriptions of varying abstraction levels) through the utilization of the provided (general, domain, structural, etc.) knowledge. Moreover, reasoning includes the case in which semantic information is further extended to identify and detect complex objects and events at an even higher level.

Two main issues related to multimedia reasoning are the representation formalism and the type of knowledge modelled. The relevant literature considers various approaches ranging from ad hoc representations to logic-based ones. Ad hoc representations seem to have been favoured during the last decades, as they are closer to human intuition and provide efficient means for structuring the required knowledge in a easy to handle, yet effective for the targeted application, way. However, the emergence of the Semantic Web and the vision of shared and interoperable metadata and semantics have affected the more recent works. With respect to the possible employed knowledge types, the use of visual and spatio-temporal information is probably the most common one. Fusing visual with auditory and/or textual information has also received quite strong interest, as the richness of multimedia data lies in their multimodal nature.

A representative list of annotation approaches utilizing explicit knowledge includes the following:

1. The ontology-based approaches presented in [349], [338], for the recognition of complex objects, where three distributed knowledge-based systems drive the image processing, the mapping of numerical data into symbolical data and the semantic interpretation.
2. The enhanced by user-defined rules ontology-based system for fuel and pancreatic cell images annotation [342], [350], [351].The rules determine the mapping between the low-level features and the respective domain concepts.
3. The ontology-based video annotation approach of [341], where rules are used to determine the algorithms and execution order for the detection of the supported domain concepts.
4. The work in [352], where a DL-based reasoner with a pseudo-extension to provide support uncertainty handling is used to infer semantic descriptions based on learned domain concept definitions.
5. The DL-based approaches presented in [353], [354] for acquiring scene interpretations utilizing domain knowledge at different levels.
6. The approach proposed in [355] for using DLs for the descriptpion of both the form and content of multimedia documents so that queries on both structural and conceptual similarity are enabled.
7. The approaches towards augmenting domain definitions with visual descriptions exploiting the WordNet corpus of [357], [356]
8. The rule-based approaches presented in [358], [359]

Figure 4-3 presents the architecture for the knowledge-assisted analysis system. According to the framework, the knowledge-assisted analysis is decomposed in three main phases:

1. Classification Pre-processing
2. Classification
3. Classification Post-Processing

During the classification pre-processing, an image is segmented into regions, and subsequently, low level visual features are extracted. In the second phase, the initial classification takes place by assigning a set of semantic label and confidence value pairs to each segment. In the framework, multiple classification algorithms could be applied in parallel, and their results are fused. The final step corrects and enhances the results of the previous phases. More specifically, refinement of the initial labels and the segmentation may occur. The former is done by applying a genetic algorithm that uses

spatial knowledge, whereas the latter is achieved through a novel semantic approach to image segmentation.

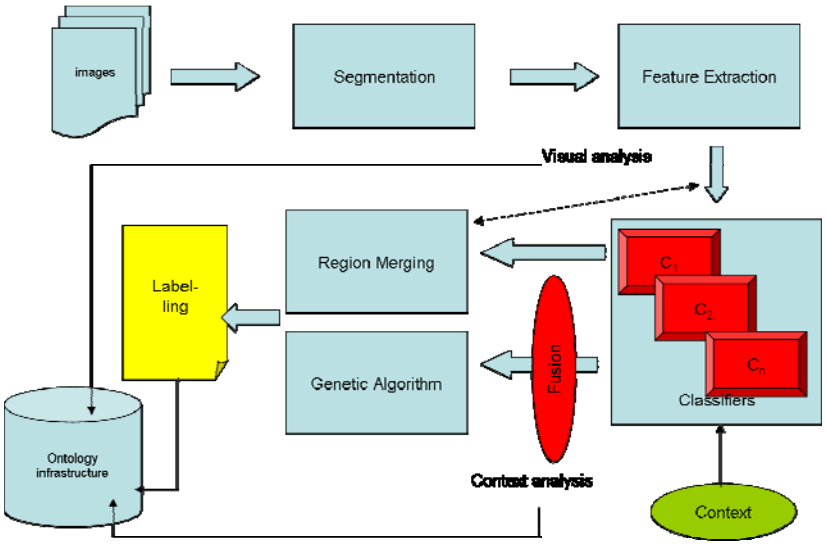

Figure 4-3: The architecture of the K-Space KAA system

The segmentation algorithm employed in the initial implementation follows the RSST approach. The descriptors to be extracted are those defined in MPEG-7, and include:

- Scalable Colour (SC),
- Homogeneous Texture (HT),
- Region Shape (RS),
- Edge Histogram (EH),
- Color Structure (CS), and
- Color Layout (CL).

The output of the descriptors complies with a common XML schema and is fed to the classification modules. In this way, the development of classifiers can occur independently of the development of feature extractors and addition of further descriptors is straightforward.

a. Classification

This is the first main extensible part of the architecture. Classification techniques are decoupled from each other. Two software modules correspond to each technique: one for offline training based upon a-priori knowledge, and one for runtime region annotation. The classifiers in the runtime classification mode process the extracted descriptors of the previous step. They conform to a generic interface, according to which: (i) each classifier is exposed as a black box executable; (ii) each classifier accepts as a parameter the image to be annotated and the name of a configuration file. In the training mode, the classifiers accept as parameters the folder of the training image set, the name of the configuration file and the name of the XML containing the ground truth data (which is compliant with a generic schema). The output is also in XML according to a common schema. A different XML file is produced for each classifier-image pair. This file contains a list of possible semantic labels along with the corresponding degrees of confidence.

b. Classifier Fusion and Other Issues

The aim of the last step in the KAA process is twofold: Firstly, to fuse the output of the multiple classifiers; and secondly, to refine intermediate results and correct sub-optimal initial decisions. The refinement process follows two directions. In the first direction, spatial knowledge is employed, so that the initial semantic labelling is refined through the application of a Genetic Algorithm. Such refinement may lead to changes in the order of the semantic labels, and as such, it may produce a

modified labelling. Complementarily, initial oversegmentation of images needs to be corrected, so that adjacent segments belonging to the same semantic entity are merged.

Research focus in Papyrus

Research in this activity will focus on developing intelligent techniques for creating multiple concept representation. This activity in collaboration with WP3 will exploit the automatic metadata extracted from Papyrus content.

4.5. Multi-modal Analysis

In the literature, most of the state of the art event detection frameworks were conducted towards the video with loose structures or without story units, such as sports video, surveillance, videos or medical videos [360]. In contrast, the concept – extraction schemes were largely carried out on the news videos which have content structures. Most of the studies for extracting concept/event are conducted in a two-stage processing [361]. The first stage includes video content processing where the video clip is segmented into certain analysis units and their representative features are extracted. The second stage is called the decision – making process that extracts the semantic index from the feature descriptors to improve the framework robustness.

Recently sports video analysis, especially sports event detection has received a great deal of attention [362], [363] owing to the great commercial potentials. For video content processing, many earlier studies adopted unimodal approaches that studied the respective role of visual, audio and texture mode in the corresponding domain. However multimodal approach attracts growing attention as it captures the video content in a more comprehensive manner. In [362] a multimodal framework using combined audio, visual and textual features was proposed. A maximum entropy method was proposed in [364] to integrate image and audio cues to extract highlights from baseball video. News videos are another video source which receives great attentions from the research community. News has a rather definite structure [365] which has been exploited for content analysis [366]. Especially the idea of defining a set of semantic concepts for which detectors could be built ahead of search time has generated great interests to the researchers, including TRECVid participants. In terms of media – based features, multimodal approaches are widely adopted [367] which explore visual, audio and automatic speech recognition (ASR) transcript based features and metadata.

In the decision making stage, data mining has been increasingly adopted. For instance [368] proposed a hybrid classification method called CBROA which integrated the decision tree and association rule mining methods in an adaptive manner. However, its performance is restricted by a segmentation process and a pre-defined confidence threshold. In [369], the authors present a pioneering work on the introduction of principle component analysis (PCA) to the face recognition domain and have popularized the use of PCA for supervised classification in this domain. As far as video semantic analysis is concerned, SVM is a well – known algorithm adopted for event detection [370], in sports videos and concept extraction [371], [372] for TRECVid videos. Although SVM presents a promising generalisation performance, its training process does not scale well as the size of the training data increased [373]. On the other hand, C4.5 [374] below is a well matured representative data mining method, which was also applied for video analysis.

In [375], the authors aim at automating the video event/concept detection procedure via the combination of distance – based and rule – based data mining techniques. In particular, the authors extend the performance of RSPM algorithm [376] for rough classification including noise/outlier filtering and feature combination and selection. Then, the well known rule based algorithm C4.5 decision tree [374] is employed for further classification. In essence, one of the unique characteristics of the proposed framework is its capability of addressing the rare event/concept detection and semantic gap issues without relying on the artifacts or domain knowledge. In terms of feature extraction, multimodal features (visual and audio) are extracted for each shot based on the detected shot boundaries. Totally five visual features are extracted for each shot namely, pixel_change, histo_change, back_ground_mean, background_var and dominant_color_ratio. Here pixel_change denotes the average percentage of the changed pixels between the consecutive frames within a shot and histo_change represents the mean value of the frame-to-frame histogram differences in a shot. Another visual feature is the dominant_color-ratio [376] that represents the ratio of dominant color in the frame based on histogram analysis and is widely used for shot classification. Then region level analysis is conducted based on segmentation results. The representations of the audio features in this framework are exploited in both time-domain and frequency-domain, dividing into three distinct groups (volume related features, energy related features and spectrum flux related features). Totally, ten generic audio features are utilized. Here volume is an indication of the loudness of sound; short time energy means the average waveform amplitude defined over a specific time window; and

Spectral Flux is defined as the 2-norm of the frame to frame spectral amplitude difference vector. For each category, some statistical attributes such as mean and standard deviation are captured as the corresponding features. In addition, due to the reason that the audio track can be continuous even around the shot boundary, the volume statistics information (mean and max) is captured for the duration of 3s around the shot boundary to explore the audio track information.

The C4.5 decision tree is a classifier in the form of a tree structure where each node is either a leaf node, indicating the value of the target class from observations, or a decision node, which specifies certain tests to be carried out on a single attribute-value and which is proceeded by either a branch or a sub-tree for each of the possible outcomes of the test. Its main classification procedure is first to construct a model for each class in terms of the attribute-value pairs and thus use the induced model to categorise any incoming testing instance. The construction of a decision tree is performed though the so-called "divide and conquer" approach, ie. Recursively partition the training set with respect to certain criteria until all the instances in a partition have the same class label, or no more attributes can be used for further partitioning. The derived model summarises all given information from training data set but express it in a more concise and perspicuous manner. The testing process of the decision tree is in the form of traversing a path in a built tree from the root to a certain leaf node and the corresponding class label is assigned to the instance when it reaches a leaf node. In the proposed framework [375], given the resulting training data set from the distance based data mining process, the C4.5 decision tree algorithm is adopted to learn a classifier and the induced classification rules are represented in the form of a decision tree. In the constructed tree, each data entry consists of audio and visual features as well as the class label. The multimodal features are extracted in the feature extraction and for each shot a class label of "yes" or "no" is assigned, showing whether there is an interested event/concept or not.

In [377], the authors argue that a key to the success of any multimedia content analysis algorithm is the type of AV features employed for the analysis. These features must be able to discriminate among different target scenes classes. Many features have been proposed for this purpose. Some of them are designed for specific tasks, while others are more general and can be useful for a variety of applications. A detailed discussion on the audio-visual features is presented in [377]. In the reminder of this section, we will focus on the correlation between Audio and Visual Features and Feature space reduction. Given, an almost endless list of audio and visual features that one can come up with, a natural question to ask is whether they provide independent information about the scene content and if not how to derive a reduced set of features that can best serve the purpose. One way to measure the correlation among features within the same modality and across different modalities is by computing the covariance matrix as presented in Equation 4.5.1.

$$C = \frac{1}{n}\sum_{x \in X}(x-m)(x-m)^T, with\, m = \frac{1}{n}\sum_{x \in X} x, \tag{4.5.1}$$

Where $x = (x_1, x_2, ..., x_K)^T$ is a $K-$dimensional feature vector, χ is the set containing all feature vectors derived from training sequences, and N is the total number of feature vectors in χ. The normalised correlation between features i and j is defined by Equation 4.5.2.

$$\tilde{C}(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}} \tag{4.5.2}$$

Where $C(i,j)$ is the $(i,j)th$ element in C. The authors apply the proposed method on a set of videos containing five types of TV programs: commercials, news, live basket ball games, live football games and weather forecast. A total of 28 features are considered: 14 audio features, eight color features and six motion features. The application of the proposed approach includes hierarchical video segmentation, video shot detection and classification using HMM, scene content classification and audio content classification.

How to combine audio and visual information belongs to the general problem of multiple evidence fusion. Until now, most work in this area was quite heuristic and application-domain dependent. A challenging task is to develop some theoretical framework for joint AV processing and more generally

for multimodal processing. For example, one may look into theories and techniques developed for sensor fusion such as Dempster Shafer theory of evidence, information theory regarding mutual information from multiple sources and Baysian theory. Another potential direction is to explore the analogy between natural language understanding and multimedia content understanding. The ultimate goals of these two problems are quite similar: being able to meaningfully partition, index, summarise and categorise a document. A major difference between a text medium and a multimedia document is that, for the same semantic concept, there are relatively few text expressions, whereas there could be infinitely many AV renditions. This makes the latter problem more complex and dynamic. On the other hand, the multiple cues present in a multimedia document may make it easier to derive the semantic content.

Research focus in Papyrus

In Papyrus, this activity will advance the state-of-the art in multi-modal analysis by exploiting the effective and efficient combination of audio-visual features extracted from the digital media and also will exploit the interaction of multi-modal nature of content. This activity will exploit available low-level analysis and ontology – aided content analysis techniques.

4.6. Semantic Annotation of Textual Documents

Semantic annotation of a Web or other textual document intends to assign explicit real-world meaning to its fragments. This annotation is normally based on a semantic model, a.k.a. ontology that describes a domain such as Tourism, Medicine or Commerce. Using semantic metadata associated with text, both humans and softwares are capable of associating meaning with documents (and/or fragments). The metadata can be stored separately from the information resource it describes (stand-off annotation), or can be embedded within the resource (in-line annotation). In Section 4.1.2.3, we presented the MPEG – 7 annotation of the textual content while in this section, we present the advances in formal annotation using domain Ontology.

Semantic descriptions encapsulating semantics along with each Web document will play crucial role in formation of the next generation Web caused by the Web expansion and the increasing use of electronic publishing in many areas of human activity. At the same time, it is important to lower the cost and complexity of such tools, in order to make them applicable ubiquitously.

We may also distinguish between different types of input documents for a text annotation system:

- Structured text: this type of text normally is characterized by rigorously predefined structure, controlled vocabulary, and restricted phrasing. Example of a structured document can be automatically generated tables or the data provided in XML format.
- Semi-structured text: it is distinguished by relatively narrow content and small unambiguous vocabulary, the phrase composition may not be strictly prescribed, but certain assumptions about the document structure can be made. For example, academic papers and enterprise reports are a kind of semi-structured documents.
- Unstructured free text: in the case of free text no particular limitations on the document's vocabulary, sentence structure or text arrangement is imposed. The content can only be limited by the domain in the broad sense. The system has to face full expressiveness, ambiguity and complexity of natural language.

Usually, each listed type of input requires a different level of text analysis. Relatively simple layout-based methods, such as wrappers, can successfully deal with arbitrary types of structured documents, reaching almost perfect results. Whereas for unstructured free text documents, more sophisticated analysis methods are used.

Early attempts of semantic annotation of a Web document was achieved with SHOE system (Simple HTML Ontology Extensions), referred to as the SHOE Knowledge Annotator [378], developed at the University of Maryland. The system is based on the SHOE language, a small extension of HTML which allows Web page authors to manually annotate their Web documents with machine-readable knowledge using lists of concepts retrieved from available ontologies.

Ontobroker [379] is one of the pioneering tools providing facilities to annotate documents by using an ontology based approach. Ontobroker uses formal ontologies to extract, information from document sources and generates metadata to help with annotating documents manually. This system helps users to annotate documents by choosing concepts from ontologies. Ontobroker has been used as a baseline tool in the framework of other projects, such as OntoAnnotate and OntoMat Annotizer of the CREAM project [380] and (KA)2 [381].

AeroDAML and AeroSWARM are the tools released by the DAML Ubot project [382]. AeroDAML is a client interface while AeroSWARM is a Web service. These tools use custom or predefined ontologies and by applying natural language information extraction techniques they generate automatically DAML annotations of Web pages. AeroDAML and AeroSWARM extract such information as people, places, organization, time, nationality.

ALPHA system [383] is a prototype developed by IBM; it uses machine learning techniques to perform automatic annotation of documents. After a training phase, in which a human selects the keywords related to a set of given concepts, the system uses statistical approaches to identify concepts and their relations in the document.

Annotea [384] is a platform for semantic annotation created under the SWAD W3C consortium project. The annotations are stored on a central server, while the client interface is available as an extension of the Web browser (Annozilla) or as standalone client (Amaya). The resources can be html pages or text document. The annotation is based on RDF tags that can be stored together with the document locally or in a centralized repository. The process is manual and the user is supported in the choice of concepts from existing ontologies.

COHSE (Conceptual Open Hypermedia Services Environment) prototype [385] aims at the use of metadata to support the construction and navigation of links in the Semantic Web. COHSE relies on ontology reasoning services and a Web-based open hypermedia link service. The tool recognizes two sources of metadata: words or word combinations in the textual part of the resource itself, including XML tags, known to represent a concept; and concepts added explicitly through some metadata annotation process using ontology.

Lixto [386] is a user-friendly tool which allows to interactively generate wrapper programs. The user can convert HTML documents into XML structured by visually selecting relevant pieces of information. The system generalizes the selection and returns to the operator all currently matched instances of a pattern, i.e., concept instances, such as Price, based on these generalizations. The feedback is provided by immediately highlighting each pattern instance, this way eliminating the tedious coding, running and then debugging process in the programming alternative. Then, the user can choose an approach to improve extraction: by imposing additional conditions or further generalizations. Lixto is tolerant of content changes if they do not affect the structure of the Web page.

Armadillo [387] is a system for unsupervised automatic domain-specific annotation on large repositories. This tool employs an adaptive information extraction algorithm, learning extraction pattern from a seed set of examples and generalize them. Bootstrapping process is repeated until the user is satisfied with the quality of extracted annotations. Learning is seeded by extracting information from structured sources, such as databases and digital libraries, or a user-defined lexicon. Information from a database already wrapped is exploited in order to provide automatic annotation of examples for the other structured sources or free text. Retrieved information is then used to partially annotate new set of documents. Then, annotated documents are used to bootstrap learning, and the process iterates.

Another tool that was tested on a large scale documents is SCORE (the Semantic Content Organization and Retrieval Engine) [388]. The tool integrates several information extraction methods, including probabilistic, learning, and knowledge-based techniques, and then combines the results from different classifiers. SCORE uses an ontology divided into two parts: the WorldModel, considered as a definitional component, and the Knowledgebase, representing a subset of the real world and used as an assertional component. The definition of both components is delegated to domain experts. The extraction process is driven by the WorldModel and is governed by a set of extraction and enhancement rules. The results are better if the source is more structured. In SCORE named entities detection problem is approached by integrating the results of several classification algorithms: probabilistic (Bayesian), learning (Hidden Markov Models), and knowledge-based techniques. The system attaches different weights to each classifier depending on the accuracy of the algorithm. Then, basing on these rates, the results from all classifiers are combined in order to produce the final annotation. SCORE also is able to resolve ambiguities. It is fulfilled using two different methods: classification-based, which associates entities with document categories basing on hand-crafted or automatically induced rules, and knowledgebase method, which uses hierarchical relationships between the entities of the Knowledgebase. To combine both methods SCORE assigns different weights for each match depending on the context where the entity appears.

Magpie [389] is a tool assisting the interpretation of the Web pages by mapping them to ontological descriptions. It uses a parser with wrapper induction methods that annotate the part of the page according to existing ontologies.

CAFETIERE (Conceptual Annotation oF Events, Terms, Individual Entities and RElations) [390] is a tool for the semi-automatic generation of XML annotations of text documents. It involves basic linguistic pre-processing, ontology linkage, and rule-based analysis for information extraction. The system distinguishes between three types of annotations: (1) structural – head, body, paragraphs, etc.; (2)

lexical – person's names, organizations and other entity instances, and (3) semantic/conceptual – entities, relationships and events. The tool interacts with ontology by performing the lookup of existing instances of the terms of ontology and allowing to exploit concepts structure, i.e., to impose constraints on concept attributes according to their type. However, rules are hand-crafted: they must be elaborated by linguist experts together with domain experts.

SemTag [391] is an application that performs automated semantic tagging of large corpora. It is based on the Seeker platform which provides functionality for the needs of automated semantic annotation algorithms: data store, crawler, indexer, and other tools for technical support of the annotation process. SemTag annotates websites with terms from TAP ontology, using corpus statistics to improve the quality of tags. The TAP knowledge base contains lexical and taxonomic information about popular objects such as music, movies, authors, sports, autos, health, and others. SemTag detects the occurrence of these entities in the Web pages through a lookup phase. Then, it disambiguates them using a vector-space model.

KIM platform [392] provides a Knowledge and Information Management (KIM) infrastructure and services for automatic semantic annotation, indexing, and retrieval of unstructured and semi-structured content. Within the process of annotation, KIM also performs ontology population. As a baseline, KIM analyzes texts and recognizes references to named entities (persons, organizations, locations, dates).

The ontology-based method of Embley et al. Ontos [393] is intended for processing preferably semi-structured Web pages with multiple records and relies primarily on regular expressions. The annotation process [394] is organized into five steps: (1) an ontology describing an area of interest is developed; (2) this ontology is parsed to generate a database scheme and to construct rules for matching constants and keywords; (3) a record extractor is invoked to obtain data from the Web: it divides an unstructured page into individual chunks, cleans them by removing markup-language tags, and provides them as individual unstructured record documents for further processing; (4) the matching rules generated by the parser are used to recognize the objects of the ontology; (5) eventually, the database scheme is populated by using heuristics, which identify appropriate attribute-value pairs and construct database tuples.

The phenomenon of folksonomy receives a lot of attention and support. According to folksonomy, the semantic annotation is fulfilled on-line by users that can decide which tags to use. SAHA [395] is a browser-based annotation tool which can be used to annotate Web pages. With Saha, annotation process can be distributed and used without installing any annotation software on user's computer. Annotations are based on a user-defined metadata annotation schema defined in OWL. Annotations are stored in a database, from which they can be retrieved by other applications using multi facet search paradigm.

KATIA [396] is an authoring annotation system that helps users to annotate Web documents. Users can annotate documents at several granularity levels, from the entire Web site structure down to the text level. Annotation can be directly attached to documents, paragraphs, sentences, words and even letters. Every time an author makes a new document, an RDF description is automatically created to stamp the date and the author's reference. When a file is placed or moved inside the Web site, a RDF description is also created to indicate the position of this resource inside the current site. At the text level, the user can select a text or image and choose a corresponding class in the ontology editor to instantiate a new annotation.

Wrapper induction methods such as Stalker [397] and BWI [398] try to infer patterns for marking the start and end points of fields to extract. In contrast to NLP-based approaches, wrappers do not utilize any linguistic patterns, but mine the information using delimiter-based extraction rules. When the learning stage is over, complete phrases related to the target concepts are identified. The biggest advantage of wrappers is that they need small amount of training data. However, whether they are generated manually or semi-automatically, wrappers strongly rely on document structure and work best for collections of rigidly structured Web pages.

Nevertheless, the success of the listed methods is largely determined by their focus on identification and classification of various named entities such as locations and personal names. While, generally

speaking, semantic annotation is not restricted to only this type of information. In our perspective, the semantic model may contain concepts characterizing varied aspects of a domain of interest.

It is clear that manual annotation of the Web documents cannot be an affordable solution, except for small, limited applications. Annotating documents by hand is a time-consuming, error prone and expensive process, and thus cannot be applied to large collections of data. In situations, where a rich annotation schema is provided, the task becomes even more difficult because human annotators naturally cannot keep in memory all entities and inter-connections of the model. Moreover, in complex cases costly domain expertise is required to decide on the correct annotation.

NLP-based tools by their nature are very restrictive to the correctness of input text and sensitive to grammatical errors, failing to terminate or provide a parse for erroneous clauses from the viewpoint of a parsing algorithm. Such cases are very frequent on the Web, where we often find syntactically incorrect constructions. Moreover, NLP systems are normally efficient in specific domains and type of text. For example, performance of the statistical machine learning based systems highly depends on the similarity of the new data set to the training corpus. As for manually constructed general-purpose NLP systems, creation of such systems requires huge investments, long-time development, efforts of experts in various fields, and thus, is rarely realized in practice.

Semi-automatic methods is a way to realize a compromise between the amount of human participation required to assist the annotation method in getting good quality answers, on the one hand, and a degree of automation, on the other hand.

Research focus in Papyrus

Based on our previous work, called Cerno [399], the one of the research aspects in Papyrus will focus on addressing this trade-off. Cerno can be classified as an ontology-based extraction method. It fundamentally differs from NLP-based tools, as it uses a lightweight robust context-free parser in place of tokenization and part-of-speech recognition. Cerno does not have the learning phase, and instead is tuned manually when being ported to a particular application, substituting or extending domain dependent components. Moreover, it does not necessarily require a gazetteer or knowledge base of known proper entities; rather it infers their existence from their structural and vocabulary context, in the style of software analyzers. The application of the patterns to documents is similar to some of existing ontology-based methods, such as Embley's approach. While Embley's method relies primarily on regular expressions, the proposed approach combines high-speed context-free robust parsing and simple word search. The method applies a set of rules constructed beforehand that guide the annotation process, some of which are generic and can be reused in other domains. In contrast to wrapper induction approaches, the proposed method uses context-independent parsing and does not require any strict input format. This allows application of the approach to more general cases where the information is contained in arbitrary text documents. Other research aspects will include extending the state-of-the art ontology based knowledge extraction techniques.

5. Semantic Search and Ontologies

The increasing availability of semantic data mainly through the Semantic Web creates the need for semantic search engines which can take advantage of the structured information in the available ontologies to support complex user information needs.

Creating a descriptive and detailed ontology of a domain is not enough for a successful semantic web application. It is necessary to be able to make the contents of the ontology accessible to the end users. Ontologies are complex structures with specific vocabulary and, as a result, have particular needs in relation to both browsing and searching their contents.

Tools for accessing data contained in ontologies and knowledge bases are not knew, several have been implemented before using different design approaches which reach various levels of expressivity and user friendliness.

Browsing an ontology is accomplished through an appropriate visualization. This issue is presented in detail in Section 3.7.

For semantic search, there are several approaches, which may be grouped in the following main categories:

- Querying
- Form-based querying
- Natural Language
- Keyword-based

These four approaches are presented in detail in the following sections.

5.1. Querying

Most knowledge stores provide facilities for querying through the use of some formal language. These languages are homologous to the use of SQL for interrogating traditional relational databases. A detailed report presenting them may be found in [51]

Here we present the most prominent ones, namely RQL[97], SPARQL[98] or SeRQL[99].

RQL is a typed language following a functional approach and supports generalized path expressions (GPEs) featuring variables on both labels for nodes (i.e., classes) and edges (i.e., properties). RQL relies on a formal graph model (as opposed to other triple-based RDF ontology languages) that captures the RDF modeling primitives and permits the interpretation of superimposed resource descriptions by means of one or more schemas.

An example of an RQL query may be the following, which returns Museum resources and their modification date.

```
SELECT X, Y
FROM Museum{X}.{Z}last_modified{Y}
```

SPARQL is an RDF query language. SPARQL is a recursive acronym standing for SPARQL Protocol and RDF Query Language. As the name implies, SPARQL is a general term for both a protocol and a query language.

[97] http://139.91.183.30:9090/RDF/RQL/

[98] http://www.w3.org/TR/rdf-sparql-query/

[99] http://www.openrdf.org/doc/sesame/users/ch06.html

Most uses of the SPARQL acronym refer to the RDF query language. In this usage, SPARQL is a syntactically-SQL-like language for querying RDF graphs via pattern matching. The language features include basic conjunctive patterns, value filters, optional patterns, and pattern disjunction.

The SPARQL protocol is a method for remote invocation of SPARQL queries. It specifies a simple interface that can be supported via HTTP or SOAP that a client can use to issue SPARQL queries against some endpoint.

An example of a SPARQL query is the following, which returns all the capitals of Afrika:

```
PREFIX abc: <http://example.com/exampleOntology#> .
SELECT ?capital ?country
WHERE {
  ?x abc:cityname ?capital ;
     abc:isCapitalOf ?y.
  ?y abc:countryname ?country ;
     abc:isInContinent abc:Africa.
}
```

Variables in SPARQL are indicated by a "?" or "$" prefix. Bindings for ?capital and the ?country will be returned in this example. The SPARQL query processor will search for sets of triples that match these four triple patterns, binding the variables in the query to the corresponding parts of each triple. Important to note here is the "property orientation" (class matches can be conducted solely through class-attributes / properties). To make queries concise, SPARQL allows the definition of prefixes and base URIs

SeRQL ("Sesame RDF Query Language") is a RDF/RDFS query language developed as part of Sesame[100]. As an example of a query in SeRQL, the following one retrieves all painters and the paintings which they have painted, as well as the technique used in the painting

```
select Painter, Painting, Technique
from   {Painter} rdf:type {cult:Painter};
                 cult:paints {Painting} cult:technique {Technique}
using namespace
     cult = <http://www.icom.com/schema.rdf#>
```

Typing queries in formal languages provides the greatest level of expressivity and control for the user. However, these query languages have a fairly complex syntax, require a good understanding of the data schema and are error prone due to the need to type long and complicated URIs. These languages should not be seen as an end user tool. The user needs to learn the complex syntax of a formal query and also to know the underlying schema and the literals expressed in the RDF data.

[100] http://www.openrdf.org/

5.2. Form-Based Querying

One step towards a more user-friendly interface is adding support for visual formal querying to ontology editing environments. For instance, Protégé [26] provides the Query Interface (Figure 5-1), where one can specify the query by selecting some options from a given list of concepts and relations.

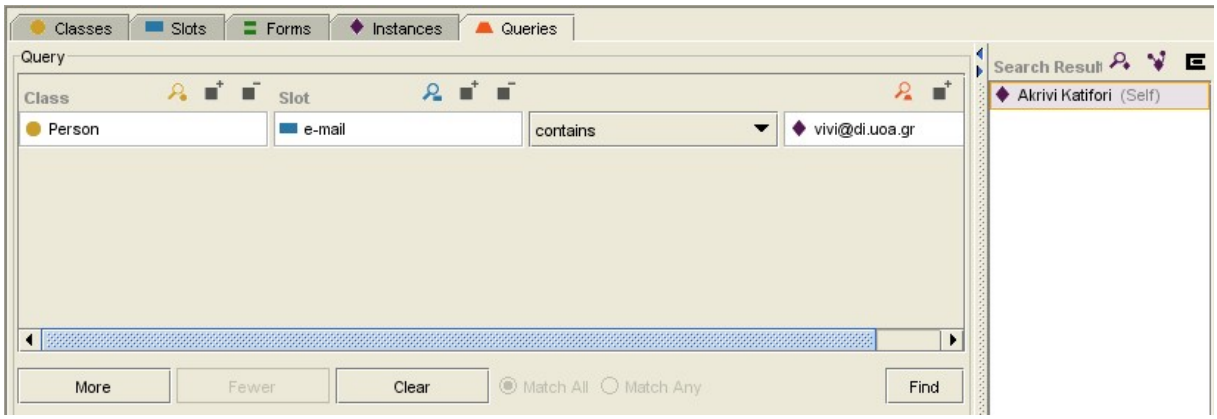

Figure 5-1. The Queries tab in Protégé

Another example of a forms-based interface is provided by the KIM [99] knowledge management platform.

5.3. Keyword-based

Probably due to the extraordinary popularity of search engines such as Google, people have come to prefer search interfaces which offer a single text input field for the user to describe his/her information need. Text interfaces have a role to play because they are familiar to end users, benefit from very good support both on the desktop and in Web interfaces, and are easily available on all types of devices. Users are very familiar with these interfaces due to their widespread usage. Compared with formal queries, keyword queries have the following advantages:

- Simple syntax: they are simply lists of keyword phrases
- Open vocabularies: the users can use their own words when expressing their information needs.

While employing this kind of interface is straightforward for full text search systems, using it for conceptual search requires an extra step that converts the user's query into semantic restrictions like those expressed in formal search languages.

Translating keywords to formal queries has been investigated both in the information retrieval and the database communities. Solutions for keyword queries over databases are presented in [88], [89], [90]. There also exist approaches that specifically deal with keyword interfaces for semantic search engines. For example, [91] [92], specifically tackle XML data by translating keyword queries to XQuery[101] expressions. However, none of them can be directly applied to semantic search on RDF data, since the underlying data model is a graph rather than relational or tree-shaped XML data.

There exist approaches that specifically deal with keyword interfaces for semantic search engines.

An example of an interface for semantic keyword search over ontologies is **SemSearch** [27], a concept-based system which aims to have a Google-like interface. It requires a list of concepts (classes or instances) as an input query. For example 'news:PhD Students' asksfor all instances of News related with PhD students. This approach allows the use of a simple text field as input but it requires a good knowledge of the domain ontology and it cannot always capture the meaning intended by the users.

This problem has been tackled recently by [93], [94]. In [94], a more generic graph-based approach has been proposed to explore the connections between nodes that correspond to keywords in the query. This way, all interpretations that can be derived from the underlying RDF graph can be computed.

However, there are remaining challenges like:

- How to deal with keyword phrases that are expressed in the user's own words which may be different from the ontology data.
- How to find the relevant query when keywords are ambiguous.
- How to return the relevant queries as quickly as possible (scalability).

5.4. Natural Language

Natural language support for querying transforms a query given in natural language into formal queries to the knowledge base. These methods are appealing to users but generally suffer from limitations in the expressiveness of the underlying supported language.

A mature example of a system employing a method of this kind is **Aqualog** [28]. It uses a controlled vocabulary for querying ontologies as well as a learning mechanism, so that its performance improves over time in response to the vocabulary used by the end users. The natural language query is then converted into a set of ontology-compatible triples that are then used to extract information from a

[101] http://www.w3.org/TR/xquery/

knowledge strore. It utilizes shallow parsing and WordNet and, as a result, it requires syntactically correct input, effective mostly for simple queries expressed as questions.

Orakel [29] is another Natural Language Inteface (NLI) to knowledge bases. It supports compositional semantic construction which helps it support questions involving quantification, conjunction and negation. However, these advanced features require a mandatory customization of the system whenever it is ported to a new application domain.

ONLI (Ontology Natural Language Interaction) [96] is a natural language question answering system used as front-end to the RACER[102] reasoner and to nRQL [30], RACER's query language. ONLI assumes that the user is familiar with the ontology domain and works by transforming the user's natural language queries into nRQL.

Querix [95] is another ontology based question answering system that translates generic natural language queries into SPARQL. In case of ambiguities, Querix relies on clarification dialogues with users.

QuestIO [97] (Question – Based Interface to Ontologies) is a NLI system for accessing structured information from a knowledge base. It is open-domain (or customizable to new domains with very little cost), with the vocabulary not being pre-defined but rather automatically derived from the data existing in the knowledge base. The system works by converting NL queries into formal queries in SeRQL (although other languages could be used).

If an NLI will be included in Papyrus has not yet been defined, as it is an issue for the user requirements. User interviews so far have shown a certain amount of scepticism on the part of the history researchers and a strong preference for keyword-based interfaces. Whether this is a result of greater familiarity with keyword-based interfaces, as opposed to NLIs, of lack of trust towards NLIs or in fact their limited usefulness towards Papyrus users, has yet to be clarified.

Lastly, it should be mentioned here that in the area of Natural Language Interfaces, there is a group of approaches that work towards a natural language representation of ontologies. **NaturalOWL**[103] **[98]** is an example of such a system. It is an open-source multilingual natural language generator that produces descriptions of instances and classes, based on a linguistically annotated ontology. It supports OWL DL ontologies with RDF linguistic annotations. It is written in Java and it is accompanied by a Protégé plug-in.

All the aforementioned methods will be investigated as to their suitability in the context of Papyrus. However, it seems that as the Papyrus user group is not comprised by otnology experts, formal query languages such as SPARQL will not be among their top choices as to usability. As a first insight to be confirmed by the detailed use requirements history researchers seem to be more confortable with keyword interfaces, even more so than NL ones. An appropriate visual browsing methods also may be very important. User requirements will show whether an advanced keyword interface accompanied by an effective visualization for browsing will be sufficient.

[102] http://www.sts.tu-harburg.de/~r.f.moeller/racer/

[103] http://www.aueb.gr/users/ion/software/NaturalOWL.tar.gz

6. Related Projects

There are several projects and synergies, either concluded or in progress, the results of which have been taken into account for the Papyrus projects. Here we present the most relevant ones of those investigated.

NeOn[104] is a project co-funded by the European Commission's Sixth Framework Programme that started in March 2006 and has a duration of 4 years. Its aim is to advance the state of the art in using ontologies for large-scale semantic applications in distributed organizations. Particularly, it aims at improving the capability to handle multiple networked ontologies that exist in a particular context, are created collaboratively, and might be highly dynamic and constantly evolving. The project is still at an early stage but there is possibility for possible interaction with PAPYRUS, as among NeOn's results there is an extensible Ontology Engineering Environment, NeOn Toolkit[105], that contains plug-ins for ontology management and visualization. NeOn toolkit has been investigated and at this moment it seems that its support for multilingual ontologies does not fatisfy Papyrus needs on this issue. Furthremore, NeON Toolkit is still under development and in our case it would be probably best to proceed with a more tested and stable solution.

The EU IST integrated project **Semantic Knowledge Technologies** (**SEKT**)[106] developed and exploited semantic knowledge technologies. Core to the SEKT project has been the creation of synergies by combining the three core research areas ontology management, machine learning and natural language processing. The project is oriented to knowledge sharing within commercial and public organizations, whereas PAPYRUS' users are a more diverse and open group with history interests. To this end the results of SEKT are not directly applicable, they would need to be adapted to accommodate multilingual ontologies with temporal features.

The **IDoRA**[107] system (Intelligent Document Retrieval and Analysis). IDoRA is currently being integrated with the news gathering system Europe Media Monitor (EMM) in order to fight the information overflow and to overcome the language barrier with the purpose of supporting the European Commission and Member State institutions. There are PAPYRUS related elements in the work performed for IDoRA and EMM, as multilingual and cross-lingual retrieval of documents and mainly news items are an important goal, as well as identification of terms. The results of these projects will be taken into account, however, it is certain that they won't be directly applicable to PAPYRUS, as PAPYRUS will be ontology-based and support search and retrieval of historical archive documents and the extraction of ontology-modeled historical information.

The MESH[108] (Multimedia Semantic Syndication for Enhanced News Services) Integrated Project aims to apply multimedia analysis and reasoning tools, network agents and content management techniques to extract, compare and combine meaning from multiple multimedia sources, and produce advanced personalized multimedia summaries, deeply linked among them and to the original sources to provide end users with an easy-to-use "*multimedia mesh*" concept, with enhanced navigation aids. A step further will empower users with the means to reuse available content by offering media enrichment and semantic mixing of both personal and network content, as well as automatic creation from semantic descriptions. Encompassing all the system, dynamic usage management will be included to facilitate agreement between content chain players (content providers, service providers and users). In a sentence, the project will create multimedia content brokers acting on behalf of users to acquire, process, create and present multimedia information personalized (to user) and adapted (to usage environment). These functions will be fully exhibited in the application area of news, by

[104] http://www.neon-project.org/web-content/

[105] http://www.neon-toolkit.org/

[106] http://www.sekt-project.com/

[107] http://langtech.jrc.it/

[108] http://www.mesh-ip.eu

creation of a platform that will unify news organizations through the online retrieval, editing, authoring and publishing of news items

The **NEWS**[109] [65] project aims at providing solutions which help news agencies to overcome limitations in their current workflows and increase their productiveness and revenues. In order to reach this aim, the NEWS project makes use of Semantic Web technologies. NEWS is a research and development project funded by the European Commission under contract FP6 001906 in the framework of the Information Society Technologies (IST) programme. In that sense, the work developed in the NEWS project covers the following topics:

- Ontology development using Semantic Web standards to define ontologies for the news industry.
- Annotation implementing a semantic annotation component which automatically produces metadata annotations for news items.
- Deductive Database that allows keyword-based queries
- Entity Identification which associates NEWS Ontology instances with entities recognized in the news item by the annotation component

[109] http://www.news-project.com

7. Conclusions

This document has presented an overview of the State of the Art in all research issues related to Papyrus. All relevant aspects of ontologies, including querying over semantically structured information have been thoroughly investigated and recorded here. Furthermore, a full presentation of content analysis issues has been included, as the analysis of the News material will be essential in the context of Papyrus. More specifically, we provided a detailed survey of various off-the-shelf tools for analysis of audio, video and textual content. Apart from describing existing work in the related areas, the deliverable provides insights into specific research focus in Papyrus within each given area.

The study of the state of the art has shown that there is still work to be done in the related areas, in order to satisfy the challenges posed by the Papyrus project. A main challenge is how to model the special characteristics of an international historical ontology, namely multilinguality in combination with temporal characteristics. Another important issue is the selection of appropriate visualization and querying interfaces tailored to the user groups interested in Papyrus. This task will be the result of the identification of user needs and requirements, to be accomplished in the following months of the project.

8. References

[1] A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis, E. Giannopoulou, Ontology Visualization Methods - A Survey, ACM Computing Surveys, Volume 39 , Issue 4 (2007)

[2] H. Alani, TGVizTab: An Ontology Visualization Extension for Protégé. In Proceedings of Knowledge Capture (K-Cap'03), Workshop on Visualization Information in Knowledge Engineering, Sanibel Island, Florida, USA, 2003

[3] K. Babaria, Using Treemaps to Visualize Gene Ontologies, Human Computer Interaction Lab and Institute for Systems Research, University of Maryland, College Park, MD USA, 12/04/2004, available at www.cs.umd.edu/hcil/treemap/GeneOntologyTreemap.pdf, 2004

[4] E. H. Baehrecke, N. Dang, K. Babaria, B. Shneeiderman, Visualization and analysis of microarray and gene ontology data with treemaps. BMC Bioinformatics, available at http://www.biomedcentral.com/1471-2105/5/84, 2004

[5] A. Bosca, D. Bomino, P. Pellegrino, OntoSphere: more than a 3D ontology visualization tool. In Proceedings of SWAP, the 2nd Italian Semantic Web Workshop, Trento, Italy, December 14-16, CEUR, Workshop Proceedings, ISSN 1613-0073, online http://ceur-ws.org/Vol-166/70.pdf, 2005

[6] P. Eklund, Visual Displays for Browsing RDF Documents. In Proceedings of the 7th Australasian Document Computing Symposium, Sydney, Australia, December 16, 2002

[7] P. Eklund, N. Roberts, S. P. Green, OntoRama: Browsing an RDF Ontology using a Hyperbolic-like Browser, In Proceedings of the First International Symposium on CyberWorlds (CW2002), Theory and Practices, IEEE press, 405-411, 2002

[8] J. Lamping, R. Rao, The Hyperbolic Browser: A Focus + Context technique for Visualizing Large Hierarchies. Journal of Visual Languages and Computing, vol. 7, 33-55, 1996

[9] J. S. M. Lee, G. Katari, R. Sachidanandam, GObar: A Gene Ontology based analysis and visualization tool for gene sets, BMC Bioinformatics, 2005

[10] N. F. Noy, R. W. Fergerson, M. A. Musen, The knowledge model of Protege-2000: Combining interoperability and flexibility. In Proceedings of 2nd International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000), Juan-les-Pins, France, 2000

[11] B. Parsia, T. Wang, J. Goldbeck, Visualizing Web Ontologies with CropCircles, In Proceedings of the 4th International Semantic Web Conference, November 6 -10, 2005

[12] B. Shneiderman, Tree visualization with Tree-maps. A 2-d space-filling approach. ACM Transactions on Graphics. Vol. 11, No. 1, September 1992, 92-99, 1992

[13] M. Sintek, Ontoviz tab: Visualizing Protégé ontologies, http://protege.stanford.edu/plugins/ontoviz/ontoviz.html, 2003

[14] K. X S. Souza, A. D. Dos Santos, S. R. M. Evangeista, Visualization of Ontologies through Hypertrees. In Proceedings of the Latin American conference on Human-computer interaction, Rio de Janeiro, Brazil, 251 – 255, 2003

[15] M.-A. Storey, M. Mussen, J. Silva, C. Best, N. Ernst, R. Fergerson, N. Noy, 2001. Jambalaya: Interactive visualization to enhance ontology authoring and knowledge acquisition in Protégé. In Proceedings of Workshop on Interactive Tools for Knowledge Capture, K-CAP-2001, Victoria, BC, Canada, http://www.thechiselgroup.org/jambalaya

[16] B. Suh, B. B. Bederson, OZONE: A Zoomable Interface for Navigating Ontology Information. In Proceedings of Advanced Visual Interfaces, ACM, 2002

[17] Y. Sure, J. Angele, S. Staab, OntoEdit: Guiding Ontology Development by Methodology and Inferencing. In Proceedings of International Conference on Ontologies, Databases and Applications of Semantics (ODBASE'02), Irvine, USA, 2002

[18] T. Wang, B. Parsia, Cropcircles: topology sensitive visualization of owl class hierarchies, in Proceedings of International Semantic Web Conference (ISWC 06), http://www.mindswap.org/papers/2006/cropcircles-iswc.pdf, 2006

[19] J. Wu, M.-A. Storey, A multi-perspective software visualization environment, In Proceedings of the 2000 conference of the Centre for Advanced Studies on Collaborative research, ACM, 2000

[20] S. Zhong, F. Storch, O. Lipan, M. J. Kao, C. Weitz., W. H. Wong, GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. Applied Bioinformatics 2004, 3(4): 1-5, 2004

[21] S. Zhong, L. Tian, C. Li, K. F. Storch, W. H. Wong, Comparative Analysis of Gene Sets in the Gene Ontology Space under the Multiple Hypothesis Testing Framework, In Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference, 2004

[22] A. Katifori, E. Torou, C. Vassilakis, G. Lepouras, C. Halatsis, I. Daradimos, Historical Archive Ontologies – Requirements, Modelling and Visualization, in Proceedings of RCIS 07, April 23-26, Ouarzazate, Morroco.

[23] A. Katifori, C. Vassilakis, G. Lepouras, I. Daradimos, C. Halatsis, Visualizing a Temporally-Enhanced Ontology, proceedings of the ACM AVI 06 Conference

[24] C. Vassilakis, A. Katifori, E. Torou, TimeViz: A temporal Ontology Visualization Plugin, http://oceanis.mm.di.uoa.gr/pened/index.php?c=publications#plugins

[25] C. Vassilakis, G. Lepouras, A. Katifori, t-Protégé – A Temporal Extension for Protégé, Technical Report TR-SSDBL-06-001, June 2006, available through http://t-protege.uop.gr

[26] N. Noy, M. Sintek, S. Decker, R. Fergerson, M. Musen: Creating Semantic Web Contents with Protégé-2000. IEEE Intelligent Systems 16(2) (2001) 60-71

[27] Y. Lei, V. S. Uren, E. Motta, SemSearch: a search engine for the semantic web. In Proceedings EKAW 2006, Managing Knowledge in a World of Networks, pages pp. 238-245, 2006

[28] V. Lopez, E. Motta, Ontology-driven Question Answering in AquaLog. In Proceedings 9th international conference on applications of natural language to information systems, Manchester, 2004

[29] P. Cimiano, P. Haase, P., J. Heizmann, Porting natural language interfaces between domains: an experimental user study with the ORAKEL system. In Intelligent User Interfaces pp. 180-189, 2007

[30] V. Haarslev, R. Möller, M. Wessel, Querying the Semantic Web with Racer + nRQL, In Proceedings of the Intern. Workshop on Applications of Description Logics, 2004

[31] S. Mithun, L. Kosseim, V. Haarslev, Resolving Quantifier and Number Restriction to Question OWL Ontologies, In Proceedings of the IEEE Third International Conference on Semantics, Knowledge and Grid, 29-31 Oct. 2007, pp.218-223

[32] M. Volkel, W. Winkler, Y. Sure, S. R. Kruk, M. Synak, SemVersion: A Versioning System for RDF and Ontologies, ESWC, 2005

[33] N. F. Noy, S. Kunnatur, M. Klein, M. A. Musen, Tracking Changes During Ontology Evolution, The Semantic Web – ISWC 2004, Springer Berlin / Heidelberg, 2004

[34] D. Allemang, I. Polikoff, R. Hodgson, P. Keller, J. Duley, J. and P. Chang, COVE – Collaborative Ontology Visualization and Evolution, IEEE Aerospace Conference 2005

[35] T. Kauppinen, E. Hyvönen, Bridging the Semantic Gap between Ontology Versions, Proceedings of the 11th Finnish AI Conference, Web Intelligence Symposium, Conference Series - No 20, vol. 2, pp. 63-72

[36] T. Sindt, Formal Operations for Ontology Evolution, Proceedings of the International Conference on Emerging Technologies, Minneapolis, Minnesota, August 25 – 26, 2003

[37] P. Ceravolo, A. Corallo, G. Elia, A. Zilli, Managing Ontology Evolution Via Relational Constraints, Paolo Ceravolo, Angelo Corallo, Gianluca Elia, Antonio Zilli, Knowledge-Based Intelligent Information and Engineering Systems, Springer Berlin / Heidelberg, 2004

[38] L. Stojanovic, B. Motik, Ontology evolution within ontology editors, In Proceedings of the OntoWeb-SIG3 Workshop at the 13th International Conference on Knowledge Engineering and Knowledge Management EKAW 2002, volume 62

[39] L. Stojanovic, A. Maedche, B. Motik, N. Stojanovic, User-driven ontology evolution management, Proceedings of the 13th European Conference on Knowledge Engineering and Knowledge Management EKAW, Madrid, Spain, 2002, pp. 285-300

[40] A. Maedche, L. Stojanovic, R. Studer, R. Volz, Managing multiple ontologies and ontology evolution in OntoLogging, Proceedings of the Conference on Intelligent Information Processing, Montreal, Canada, 2002, pp. 51-63

[41] P. Plessers, O. De Troyer, S. Casteleyn, Understanding ontology evolution: A change detection approach, Web Semantics: Science, Services and Agents on the World Wide Web, Volume 5, Issue 1 (March 2007), pp 39-49

[42] J. Heflin, J. A. Hendler, Dynamic ontologies on the web, In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, pp 443–449. AAAI Press / The MIT Press, 2000.

[43] M. Klein, D. Fensel, Ontology versioning for the Semantic Web. In Proceedings of the First International Semantic Web Working Symposium (SWWS), pp. 75–91, Stanford University, California, USA, July 30 – August 1, 2001.

[44] M. Klein, A. Kiryakov, D. Ognyanov, D. Fensel, Finding and characterizing changes in ontologies. In Proceedings of the 21st International Conference on Conceptual Modeling (ER2002), number 2503 in LNCS, pp. 79–89, Tampere, Finland, October 7–11, 2002

[45] D. Steven, J. Perrin, PROMPT-Viz: Ontology Version Comparison Visualizations with Treemaps. Master Of Science Thesis in the Department of Computer Science, University of Victoria, Retrieved from http://www.cs.uvic.ca/~chisel/thesis/David_Perrin_Thesis.pdf, 2004

[46] T. Kauppinen, E. Hyvönen, Modeling coverage between geospatial resources. In Posters and Demos at the 2nd European Semantic Web Conference ESWC2005, pages 49–50, Heraklion, Crete, 2005.

[47] T. Kauppinen, E. Hyvönen, Modeling and Reasoning about Changes in Ontology Time Series, chapter 11 in book: Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems. Integrated Series in Information Systems, Volume 14. Springer-Verlag, 2006.

[48] T. Kauppinen, R. Henriksson, J. Väätäinen, C. Deichstetter, E. Hyvönen, Ontology-based Modeling and Visualization of Cultural Spatio-temporal Knowledge. Proceedings of the 12th Finnish Artificial Intelligence Conference STeP 2006, Espoo, Finland, October 26-27, 2006

[49] N. F. Noy, A. Chugh, W. Liu, M. A. Musen, A Framework for Ontology Evolution in Collaborative Environments, International Semantic Web Conference - ISWC 2006

[50] V. K. Chaudhri, A. Farquhar, R. Fikes, P. D. Karp, J. Rice, OKBC: A programmatic foundation for knowledge base interoperability. In AAAI/IAAI, pp. 600–607, 1998

[51] A. Magkanaraki, G. Karvounarakis, T. Anh, V. Christophides, D. Plexousakis, Ontology Storage and Querying, Technical Report No 308, Foundation for Research and Technology, Hellas Institute of Computer Science, Information Systems Laboratory, April 2002

[52] Z. J. Zhang, Ontology Query Languages For The Semantic Web: A Performance Evaluation, Thesis for the Master of Science, University of Georgia, Athens, Georgia, Available at http://www.cs.uga.edu/~jam/home/theses/zhijun_thesis/final/zhang_zhijun_200508_ms.pdf, 2005

[53] O. Corcho, A. Gòmez – Pérez, A roadmap to ontology specification languages, Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management,Pages: 80 – 96, 2000

[54] Y. B. Guo, Z. X. Pan, J. Heflin, An Evaluation of Knowledge Base Systems for Large OWL Datasets, in Proceedings of the Third International Semantic Web Conference, Hiroshima, Japan, November 7-11, 2004

[55] J. Broekstra, A. Kampman, Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In Proceedings of ISWC 2002

[56] M. Denny, Ontology Building: A Survey of Editing Tools, November 2002, available in http://www.xml.com/pub/a/2002/11/06/ontologies.html

[57] M. Denny, Ontology Tools Survey: Revisited, July 2004, available in http://www.xml.com/pub/a/2004/07/14/onto.html

[58] J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, K. Wilkinson, The Jena Semantic Web Platform: Architecture and design, Technical report, Hewlett Packard Laboratories, 2003.

[59] B. McBride, Jena: Implementing the RDF model and syntax specification. In S. Decker et al., editors, Second International Workshop on the Semantic Web, Hong Kong, May 2001.

[60] C. Murray N. Alexander S. Das, G. Eadon, S. Ravada, Oracle Spatial Resource Description Framework (RDF), 10g Release 2 (10.2), 2005.

[61] K. Wilkinson, C. Sayers, H. Kuno, D. Reynolds, Efficient RDF Storage and Retrieval in Jena2, Technical report, Hewlett Packard Laboratories, 2003, HPL-2003-266

[62] N. Alexander, X. Lopez, S. Ravada, S. Stephens, J. Wang, RDF Data Model in Oracle, Oracle Corporation

[63] Integration of jena in protégé-owl, available at http://protege.stanford.edu/plugins/owl/jena-integration.html

[64] N. Fernandez - Garcia, L. Sanchez-Fernandez, Building an Ontology for NEWS Applications. In Poster Session of the 3rd International Semantic Web Conference, ISWC, 2004

[65] L. Zapf, N. Fernández-García, L. Sánchez-Fernández, The NEWS Project – Semantic Web Technologies for the News Domain, 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, 30.November-1.December 2005, London, UK

[66] N. Guarino, Formal Ontology and Information Systems, 1st International Conference on Formal Ontology in Information Systems (FOIS'98), 3-15, Trento, June 1998.

[67] DOLCE: a Descriptive Ontology for Linguistic and Cognitive Engineering. http://www.loa-cnr.it/DOLCE.html

[68] C. Fellbaum (ed.). WordNet: An Electronic Lexical Database. MIT Press, May 1998.

[69] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, Wonderweb Deliverable D18 – Ontology Library", 2003.

[70] C. Tsinaraki, P. Polydoros, S. Christodoulakis, Integration of OWL ontologies in MPEG-7 and TVAnytime compliant Semantic Indexing, 16th International Conference on Advanced Information Systems Engineering (CAiSE'04), Riga, June 2004.

[71] R. Wille, Concept lattices and conceptual knowledge systems. Computers and Mathematics with Applications 23:493–515, 1992

[72] J. Kang and J. F. Naughton., On Schema Matching with Opaque Column Names and Data Values. In ACM SIGMOD Conference, pages 205–216, 2003.

[73] A. Doan, P. Domingos, and A. Halevy. Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach. In ACM SIGMOD Conference, pages 509–520, 2001.

[74] F. Naumann, C.T. Ho, X. Tian, L. M. Haas, and N. Megiddo. Attribute Classification Using Feature Analysis. In Proceedings of International Conference on Data Engineering (ICDE), page 271, 2002.

[75] B. He and K. C. Chang. Statistical Schema Matching across Web Query Interfaces. In ACM SIGMOD Conference, pages 217–228, 2003.

[76] H. H. Do and E. Rahm. COMA - A System for Flexible Combination of Schema Matching Approaches. In Proceedings of the International Conference on Very Large Data Bases (VLDB), pages 610–621, 2002.

[77] J. Madhavan, P. Bernstein, and E. Rahm. Generic Schema Matching with Cupid. In Proceedings of the International Conference on Very Large Data Bases (VLDB), pages 49–58, 2001.

[78] S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity Flooding: A Versatile GraphMatching Algorithm and Its Application to Schema Matching. In Proceedings of International Conference on Data Engineering (ICDE), pages 117–128, 2002.

[79] E. Rahm and P. A. Bernstein. A Survey of Approaches to Automatic Schema Matching. International Journal on Very Large Data Bases, 10(4):334–350, 2001.

[80] Shvaiko, J. Euzenat: "A Survey of Schema-based Matching Approaches". In Journal on Data Semantics, IV: 146-171, 2005.

[81] L. Popa, M. Hernandez, Y. Velegrakis, R. J. Miller, and R. Fagin. Mapping XML and Relational Schemas with Clio. In ICDE, pages 498-499, 2002.

[82] L. Popa, Y. Velegrakis, R. J. Miller, M. A. Hernandez, and R. Fagin. Translating Web Data. In VLDB, pages 598–609, 2002

[83] R. Fagin, M. Hernandez, R. J. Miller, L. Popa and Y. Velegrakis: "System and method for translating data from a source schema to a target schema", IBM Patent ARC920030001US1, Filed Mar. 2003.

[84] An, Y., Borgida, A., Mylopoulos, J., "Discovering the Semantics of Relational Tables Through Mappings", *Journal of Data Semantics VII*, 1-32, 2006.

[85] Noy, N. F. and Musen, M. A. 2001. Anchor-PROMPT: Using Non-Local Context for Semantic Matching. *Workshop on Ontologies and Information Sharing*. IJCAI, Seattle, WA, 2001

[86] Gruber, T. R. 1993. A Translation Approach to Portable Ontology Specifications, Knowledge Acquisition. Special issue: Current issues in knowledge modelling, Vol 5, Issue 2, 199-220

[87] Noy, N. F., McGuiness D. L. 2001. *Ontology Development 101: A Guide to Creating Your First Ontology*, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001

[88] Hristidis, V., Papakonstantinou, Y., DISCOVER: Keyword Search in Relational Databases. VLDB 2002

[89] Bhalotia, G., Hulgeri, A., Nakhe, C., Chakrabarti, S., Sudarshan, S., Keyword Searching and Browsing in Databases using BANKS. ICDE 2002, 431-440

[90] Balmin, A., Hristidis, V., Papakonstantinou, Y., Authority-Based Keyword Queries in Databases using ObjectRank, VLDB 2004

[91] Hristidis, V., Koudas, N., Papakonstantinou, Y., Srivastava, D., Keyword Proximity Search in XML Trees. IEEE TKDE, April 2006 (Vol. 18, No. 4) pp. 525-539

[92] Guo, L., Shao, F., Botev, C., Shanmugasundaram, J., XRANK: Ranked Keyword Search over XML Documents. SIGMOD 2003

[93] Zhou, Q., Wang, C., Xiong, M. Wang, H., Yu, Y., SPARK: Adapting Keyword Query to Semantic Search. ISWC/ASWC 2007, pp. 694-707

[94] Than, T., Cimiano, P., Rudolph, S., Studer, R., Ontology-based interpretation of keywords fro semantc search. ISWC/ASWC, 2007, pp. 523-536

[95] Kaufmann, E., Bernstein, A., Zumstein, R., Querix: A Natural Language Interface to Query Ontologies Based on Clarification Dialogs, *5th International Semantic Web Conference (ISWC 2006)*, Athens, GA, November 2006, pp. 980-981

[96] Mithun, S., Kosseim, L., Resolving quantifier and number restriction to question owl ontologies. Proceedings of the First International Workshop on question answering (QA 2007), Xian, China, 2007

[97] V. Tablan, D. Damljanovic, K. Bontcheva, A natural language query interface to structured information, to appear in the 5th European Semantic Web Conference, 2008

[98] Galanis, D., Androutsopoulos, I., Generating Multilingual Descriptions from Linguistically Annotated OWL Ontologies: the NaturalOWL System, ENLG 2007

[99] Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., Goranov, M., KIM – A semantic platform for information extraction and retrieval, Natural Language Engineering 10 (2004), 375-392

[100] M. Ghadiali, J.C.H. Poon, W.C. Siu, "Fuzzy pattern spectrum as a texture descriptor", Electronic letters, Vol. 32, No. 19, Pp.1772 - 1773, 12 Sept. 1996

[101] B.S. Manjunath, P. Salembier, T. Sikora, Introduction to MPEG-7, Multimedia Content Description Interface, John Wiley &Sons, 2003.

[102] Ho Young Lee; Ho Keun Lee; Yeong Ho Ha; Spatial color descriptor for image retrieval and video segmentation, IEEE trans. on multimedia, Vol.5, No. 3, Sept. 2003 Pp.358 - 367

[103] B.S. Manjunath, J.R. Ohm, V.V. Vasudevan, and A. Yamada, "Color and Texture Descriptors. In Special Issue on MPEG-7", IEEE Transactions on Circuits and Systems for Video Technology, 11/6 Pp. 703-715, June 2001.

[104] S.-F. Chang, T.Sikora, and A. Puri, "Overview of MPEG-7 Standard", IEEE Trans. Circuits Syst. Video Technol., vol. 11, pp. 688-695, June 2001.

[105] Y.M. Ro, H.K. Kang, "Hierarchical rotational invariant similarity measurement for MPEG-7 homogeneous texture descriptor", Electronic Letters, Vol. 36, No. 15, Pp.1268 - 1270, 20 July 2000

[106] S. Jeannin, and A. Divakaran, "MPEG-7 visual motion descriptors", IEEE trans. on circuits and systems for video technology, Vol. 11, No. 6, Pp.720 - 724, June 2001

[107] Y. Deng, B.S. Manjunath, C. Kenney, M.S. Moore, H. Shin, "An efficient color representation for image retrieval", IEEE trans. on Image processing, Vol. 10, No. 1, Pp.140 - 147, Jan. 2001

[108] B.S. Manjunath and W.Y. Ma, "Texture Features for Browsing and Retrieval of Image Data", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.18, Pp. 837-842, 1996

[109] S. Nepal, U. Srinivasan, and G. Reynolds. Automatic detection of goal segments in basketball videos. In Proc. of ACM Multimedia, pages 261-269, 2001.

[110] A.V. Ratnaike, B. Srinivasan, S. Nepal, "Making Sense Of Video Content," proc. 11th ACM international conference on Multimedia (MM'03), pp. 650-651, Berkeley, CA, USA, 2003.

[111] J. S. Boreczky and Lawrence A. Rowe. Storage and retrieval for image and video databases (spie). In Comparison of Video Shots Boundary Detection Techniques, 1996.

[112] P. Browne, A. Smeaton, N. Murphy, N. O'Connor, S. Marlow, and C. Berrut. Evaluating and combining digital video shot boundary detection algorithms. In Irish Machine Vision and Image Processing Conference, 2002.

[113] A. Smeaton, W. Kraaij, and P. Over. The trec video retrieval evaluation (trecvid): A case study and status report. In Riao 2004, 2004.

[114] C. Petersohn. Franhoffer hhi at trecvid 2004: Shot boundary detection system. In Proceedings of TRECVID 2004, 2004.
[115] H.E. Williams T. Volkmer, S.M.M. Tahaghoghi. Rmit university at trecvid 2004. In Proceedings of TRECVID 2004, 2004.
[116] J. Vermaak, P. Perex, M. Ganget, and A. Blake. Rapid summarisation and browsing of video sequences. In Proceedings of the British Machine Vision Conference, 2002.
[117] M. Cooper and J. Foote. Discriminative techniques for keyframe selection. In IEEE International Conference on Multimedia and Expo, 2005.
[118] A. Girgensohn, J. Boreczky, and L. Wolcox. Keyframe-based user interfaces for digital video. In IEEE Computer, volume 34(9), pages 61-67, September, 2001.
[119] M. Yeung and B.-L. Yeo. Time constrained clustering for segmentation of video into story units. In Proceedings of International Conference on Pattern Recognition, 1996.
[120] M. Yeung and B.-L. Yeo. Video visualisation for compact presentation and fast browsing of pictorial content. In IEEE Transactions on Circuits and Systems for Video Technology, pages 771-785, 1997.
[121] Y. Rui, Thomas S. Huang, and Sharad Mehrotra. Constructing table-of-content for video. In ACM Journal of Multimedia Systema, pages 359-368, 1998.
[122] J. R. Kender and Book-Lock Yeo. Video scene segmentation via continuous video coherence. In Proceedings CVPR, pages 167-393, 1998.
[123] Z. Rasheed and M. Shah. Scene detection in hollywood movies and tv shows. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003.
[124] J. Y. Zhou and W. Tavanapong. Shotweave: A shot clustering technique for story browsing for large video databases. In Proceedings of the Workshops XMLDM, MDDE, and YRWS on XML-Based Date Management and Multimedia Engineering-Revised Papers, 2002.
[125] Ba Tu Truong, Svetha Venkatesh, and Dorai Chitra. Scene extraction in motion pictures. In IEEE Transactions on Circuits and Systems for Video Technology, volume 13, pages 5-15, January 2003.
[126] T. Liu and J. R. Kender. Proceedings of the ieee workshop on content-based access on image and video libraries. In A Hidden Markov Approach to the Structure of Documentaries, 2000.
[127] H. Sundaram and Shih-Fu Chan. Determining computable scenes in films and their structures using audio-visual memory models. In ACM Multimedia 2000, 2000.
[128] Y. Cao, W. Tavanapong, Kihwan Kim, and JungHwan Oh. Audio-assisted scene segmentation for story browsing. In Proceedings of the International Conference on Image and Video Retrieval, 2003.
[129] H.i Sundaram and Shih-Fu Chan. Determining computable scenes in films and their structures using audio-visual memory models. In ACM Multimedia 2000, 2000.
[130] J. Huang, Zhu Liu, and Yeo Wang. Integration of audio and visual information for content-based video segmentation. In IEEE Int'l Conf. Image Processing, 1998.
[131] Ying Li and C.-C. Jay Kou. Video Content Analysis using Multimodal Information. Kluwer Academic Publishers, 2003.
[132] S. Boykin and Andrew Merlino. Communications of the acm. In Machine Learning of Event Segmentation for News on Demand, February, 2000.
[133] N. O'Hare, A. Smeaton, C. Czirjek, N. O'Connor, and N. Murphy. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2004, montreal, quebec. In A Generic News Story Segmentation System and its Evaluation, 2004.
[134] G. M. Quenot, Daniel Moraru, Stephane Ayache, Mbarek Charad, Mickael Guironnet, Lionel Carminati, Jerome Gensel, Dennis Pullerin, and Laurent Besacier. Clips-lis-lsr-labri experiments at trecvid 2004. In TREC Video Retrieval Conference, 2004.
[135] K. Hoashi, M. Sugano, K. Matsumoto, F. Sugaya, and Y. Nakajimi. Shot boundary determination on mpeg compressed domain and story segmentation experiments for trecvid 2004. In TREC Video Retrieval Conference, 2004.
[136] W. Kraaj and J. Arlandis. Trecvid-2004 story segmentation task: Overview. In TREC Video Retrieval Conference, 2004.
[137] A.G. Money, H. Agius, "Video Summarisation: A Conceptual Framework and Survey of the State of the Art", in Journal of Visual Communication and Image Recognition 2007, doi: 10.1016/j.jvcir.2007.04.002

[138] M. Furini, V. Ghini, "An Audio-Video Summarisation Scheme Based on Audio and Video Analysis", in Proc. IEEE Consumer Communications and Networking Conference (CCNC '06), Vol. 2, Las Vegas, NV, USA, pp. 1209 – 1213, January 2006.
[139] Chog-Wah, Yu-Fei Ma, Hong-Jiang Zhang, "Video Summarization and Scene Detection by Graph Modeling", IEEE Trans. Circuits and Systems for Video technology, Vol. 15, No. 2, pp. 296-305, Feb. 2005.
[140] H.S. Chang, S.S. Sull and S.U. Lee, "Efficient video indexing scheme for content based retrieval", IEEE Trans. Ciruits and Systems for Video technology, Vol. 9, No. 8, pp. 1269-1279, Dec. 1999
[141] D. DeMenthon, V. Kobla and D. Doermann, "Video Summarization by curve simplification", in Proc. 6th ACM Int. Conf. Multimedia, pp.211 – 218, 1998.
[142] A. Hanjalic and H.J. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster validity analysis", IEEE Trans. Circuits and Systems Video technology, Vol. 9, No. 8, pp. 1280-1289, Dec. 1999.
[143] M. M. Yeung and B. L. Yeo, "Video visualization for compact presentation and fast browsing of pictorial content", IEEE Trans. Circuits Systems Video Technology, Vol. 7, No. 5, pp. 771-785, Oct. 1997
[144] Y. H. Gond and X. Liu, "Video summarization using singular value decomposition" in Proc. Int. Conf. Computing Visual Pattern Recognition, Vol. 2, 2000, pp. 174 – 180.
[145] J. Nam, A. T. Tewfik, "Dynamic video summarization and visualization", in Proc. 7th Int. Conf. Multimedia, pp. 53-56, 1999.
[146] R. Lienhart, "Dynamic video summarization of home video", SPIE, Vol. 3972, pp. 378-389, Jan. 2000.
[147] M. A. Smith and T. Kanade, "Video Skimming and characterization through the combination of image and language understanding techniques" in Proc. Int. Conf. Computing Visual Pattern Recognition, pp. 775-781, 1997.
[148] X. Orriols and X. Binefa, "An EM algorithm for video summarization, generative model approach", in Proc. Int. Conf. Computing Vis., Vol. 2, pp. 335 – 342, 2001.
[149] Yu-Fei Ma Hong-Jiang Zhang , "A model of motion attention for video skimming", in Proc. Int. Conf. Image Process., vol. 1, pp 129 – 132, 2002.
[150] N. Vasconcelos, A. Lippman, "A spatio-temporal motion model for video summarisation", in Proc. Int. Conf. Comput. Vis. Pattern Recognition, pp. 361 – 366, 1998
[151] Y. F. Ma, L. Lu, H. J. Zhang, M. Li, "A user attention model for video summarisation", in Proc. 10th ACM Int. Conf. Multimedia, pp. 533 – 542, 2002
[152] Z. Liu and Y. Wang, Audio Indexing and Retrieval, Handbook of Video Databases Design and Applications, CRC Press, pp. 483 - 510, 2003.
[153] J. Saunders, Real-time discrimination of broadcast speech/music, Proc. ICASSP 1996.
[154] T. Kemp, M. Schmidt, M. Westpal, A. Waibel, Strategies for automatic segmentation of audio data, Proc. ICASSP – Istanbul, Turkey, 2000.
[155] S. S. Chen and P. S. Gopalakrishnan, Speaker, environment and channel change detection and clustering via the Bayesian information criterion, Proc. DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, VA, pp. 127–132, 1998.
[156] S. Pfeiffer, S. Fischer, and W. Effelsberg, Automatic Audio Content Analsyis, Proc. 4th ACM Int. Conf. Multimedia, Boston, MA, Nov. 18-22, pp. 21-30, 1996.
[157] E. Scheirer and M. Slaney, Construction and Evaluation of a Robust Multifeature, Speech/Music Discriminator, Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Proc. Pp. 1331-1334, Munich, Germany, 1997.
[158] C. Saraceno and R. Leonardi, Audio As a Support to Scene Change Detection and Characterization of Video Sequences, ICASSP-1997, Vol. 4, pp. 2597-2600, 1997.
[159] R.Y. Qiao, Mixed wideband speech and music coding using a speech/music discriminator, The 1997 IEEE TENCON Conference. Part 2 (of 2); Brisbane; Australia; 02-04 Dec. 1997. pp. 605-608, 1997
[160] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura, Video Handling with Music and Speech Detection, IEEE Multimedia Magazine, Vol. 5, pp. 17-25, July-Sept. 1998.
[161] S. Rossignol, X. Rodet, J. Soumagne, Automatic Characterisation of Musical Signals: Feature Extraction and Temporal Segmentation, Journal of New Music Research 28(4):281-295, 1999.

[162] M. Carey, E. S. Parris, H. Lloyd-Thomas, A comparison of features for speech, music discrimination Ensigma Ltd., Chepstow, Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on Publication Date: 15-19 Mar 1999 Volume: 1, 149-152 vol.1 , 1999.
[163] G. Williams, D. P. W. Ellis, Speech/music discrimination based on posterior probability features, EUROSPEECH'99, 687-690, 1999.
[164] J. Foote, Automatic Audio Segmentation using a Measure of Audio Novelty, Proceedings of IEEE International Conference on Multimedia and Expo, vol. I, pp. 452-455, 2000.
[165] K. El-Maleh, M. Klein, G. Petrucci and P. Kabal, Speech/Music Discrimination for Multimedia Applications, IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 2445--2448, Istanbul, Turkey, 2000.
[166] G. Lu, T. Hankinson, An Investigation of Automatic Audio Classification and Segmentation, WCC 2000 ICSP 2000, 21 - 25 August, pp 776 - 780, House of Electronics, Beijing China, 2000.
[167] H. Ezzaidi, J. Rouat, Speech, music and songs discrimination in the context of handsets variability, ICSLP-2002, 2013-2016, 2002.
[168] H. Harb Hadi, L. Cheb, Sound Recognition: a connectionist approach, Proceedings of the IEEE International Symposium on Signal Processing and its Applications ISSPA2003, July 1-4, Paris - France, 2003
[169] N. Casagrande, D. Eck, and B. Kégl. Frame-level audio feature extraction using AdaBoost, Proc. 6th International Conference on Music Information Retrieval (ISMIR 2005), 2005.
[170] J. E. Muñoz-Expósito, S. Garcia-Galón, Nicolás Ruiz-Reyes, Pedro Vera-Candeas, F. Rivas-Peña: Speech/Music Discrimination Using a Single Warped LPC-Based Feature. ISMIR 2005: 614-617, 2005.
[171] D. Kimber, L.Wilcox, Acoustic segmentation for audio browsers, Proc. Interface Conference (Sydney, Australia 96), 1996.
[172] T. Zhang and C.–C. J. Kuo, Hierarchical Classification of Audio Data for Archiving and Retrieving, Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Proc., pp. 3001-3004, Phoenix, March, 1999.
[173] G. Tzanetakis, P. Cook, Multi-Feature Audio Segmentation for Browsing and Annotation, Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, 1999.
[174] T. Zhang and C.-C. J. Kuo, Audio content analysis for online audio-visual data segmentation and classification, IEEE Trans. Speech Audio Process., vol. 9, no. 44, pp. 441–457, May 2001.
[175] L. Lu, H. Zhang, Content Analysis for Audio Classification and segmentation, IEEE Transaction on speech and audio processing, vol 10. no 7., 2002.
[176] A. Arias, J. Pinquier, R. André-Obrecht. Evaluation of Classification Techniques for Audio Indexing. 13th European Signal Processing Conference (EUSIPCO'2005), September 4-8, 2005. Antalya, Turkey, 2005.
[177] S. Ghaemmaghami, Audio segmentation and classification based on a selective analysis scheme in: Multimedia Modelling Conference, 2004. Proceedings. 10th International 5-7 Jan. 2004 ,42-48, 2004.
[178] N. Patel and I. Sethi, Audio Characterization for Video Indexing, Proc. SPIE in Storage and Retrieval for Still Image and Video Databases, Vol.2670, San Jose, pp. 373-384 , 1996.
[179] Y. Nakajima, M.. Yang Lu Sugano, A. Yoneyama, H. Yamagihara, A. Kurematsu, KDD R, Saitama; A fast audio classification from MPEG coded data. Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on 15-19 Mar 1999 Volume: 6, 3005-3008 vol.6 Location: Phoenix, AZ, USA, 1999.
[180] G. Tzanetakis and P. Cook, Sound Analysis Using MPEG Compressed Audio, Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Proc. ICASSP 2000, Vol II, pp. 761-764, Istanbul, Turkey, 2000.
[181] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. J. of Royal Statistical Society, 39(1):1 – 22, 1977.
[182] U. Srinivasan, S. Pfeiffer, S. Nepal, M. Lee, L. Gu, S. Barrass: A Survey of MPEG-1 Audio, Video and Semantic Analysis Techniques. Multimedia Tools Appl. 27(1): 105-141, 2005.
[183] M. A. H. Huijbregts, R. J. F. Ordelman and F. M. G. de Jong, Annotation of Heterogeneous Multimedia Content Using Automatic Speech Recognition, Proceedings of the Second

International Conference on Semantic and Digital Media Technologies, SAMT 2007, Lecture Notes in Computer Science, volume 4816, Springer Verlag, Berlin, pp. 78-90, 2007.
[184] S. Salcedo-Sanz, J. M. Leiva-Murillo, Offline Speaker Segmentation Using Genetic Algorithms and Mutual Information, Ieee Transactions On Evolutionary Computation, Vol. 10, No. 2, 2006.
[185] S. Cramoysan, B. Elliot, J. Fenn, J. Davies, A. Litan, A. Allan, K. Dulaney, E. Kolsky, D. Kraus, Hype Cycle for Enterprise Speech Technologies, 2006, Gartner 2006.
[186] X. Anguera , Robust Speaker Diarization for Meetings, Ph.D. Thesis, 2006.
[187] M. Yamaguchi, M. Yamashita, and S. Matsunaga, Spectral cross-correlation features for audio indexing of broadcast news and meetings, Proc. International Conference on Speech and Language Processing, 2005.
[188] J. Pelecanos, J. and S. Sridharan, Feature warping for robust speaker verification, ISCA Speaker Recognition Workshop odyssey, Crete, Greece, 2001.
[189] P. Ouellet, G. Boulianne, P. Kenny, Fravors of gaussian warping, Proc. International Conference on Speech and Language Processing, Lisbon, Portugal, 2005.
[190] R. Sinha, S. E. Tranter, J. J. F. Gales, P: C, Woodland, The Cambridge university march 2005 speaker diarisation system, European Conference on Speech Communication and Technology (Interspeech), Lisbon, Portugal, pp. 2437–2440, 2005.
[191] X. Zhu, C. Barras, S. Meignier and J.-L. Gauvain,, Combining speaker identification and bic for speaker diarization, Proc. International Conference on Speech and Language Processing, Lisbon, Portugal, 2005.
[192] Y. Moh, P. Nguyen, and J..C, Junqua, Towards domain independent speaker clustering, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Hong Kong, 2003.
[193] W.-H. Tsai, S.-S. Cheng, H.-M. Wang, Speaker clustering of speech utterances using a voice characteristic reference space, Proc. International Conference on Speech and Language Processing, Jeju Island, Korea, 2004.
[194] W.-H. Tsai, S.-S. Cheng, Y.-H Chao, H.-M and Wang, Clustering speech utterances by speaker using eigenvoice-motivated vector space models, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, USA, 2005.
[195] M. Collet, D. Charlet and F. Bimbot, A correlation metric for speaker tracking using anchor models, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, USA, 2005.
[196] D. Sturim, D. Reynolds, E. Singer and J.P. Campbell, Speaker indexing in large audio databases using anchor models, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Salt Lake City, USA, 2001.
[197] W. Chan, T. Lee, N. Zheng, and hua Ouyang, Use of vocal source features in speaker segmentation, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, 2006.
[198] P. Delacourt, P. and C. J. Wellekens, DISTBIC: A speaker-based segmentation for audio data indexing, Speech Communication: Special Issue in Accessing Information in Spoken Audio 32, 111–126, 2000.
[199] L. Lu, H. Zhang, Content Analysis for Audio Classification and segmentation, IEEE Transaction on speech and audio processing, vol 10. no 7. October 2002.
[200] L. Lu, and H.-J. Zhang, Speaker change detection and tracking in real-time news broadcasting analysis, ACM International Conference on Multimedia, pp. 602–610, 2002.
[201] M. Kotti, E. Benetos and C. Kotropoulos, Automatic Speaker Change Detection with the Bayesian Information Criterion using MPEG-7 Features and a Fusion Scheme, Proc. of IEEE International Symposium Circuits & Systems(ISCAS 06), 21-24 May, Island of Kos, Greece, 2006.
[202] J. Ajmera, I. McCowan, and H. Bourlard, Robust speaker change detection, IEEE Signal Process. Lett., vol. 11, no. 2, pp. 649–651, Jun. 2004.
[203] A. G. Adami, S. S. Kajarekar, and H. Hermansky, A new speaker change detection method for two-speaker segmentation, Proc. Int. Conf. Acoustics, Speech Signal Process. (ICASSP), vol. 4, Phoenix, AZ, pp. 3908–3911, 2002.
[204] J. Ferreiros-López and D. P. W. Ellis, Using acoustic condition clustering to improve acoustic change detection on broadcast news, Proc. Int. Conf. Spoken Language Process. (ICSLP), Beijing, China, pp. 568–571, 2000.

[205] L. Wilcox, F. Chen, D. Kimber, and V. Balasubramanian, Segmentation of speech using speaker identification, in IEEE Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP), vol. 1, pp. 161–164, 1994.
[206] T. Kemp, M. Schmidt, M. Westphal, A. and Waibel, Strategies for automatic segmentation of audio data, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Istanbul, Turkey, pp. 1423–1426, 2000.
[207] H. Wactlar, A. Hauptmann, A. and M. Witbrock, News on-demand experiments in speech recognition, ARPA STL Workshop, 1996.
[208] M. Nishida, T. Kawahara, Unsupervised speaker indexing using speaker model selection based on bayesian information criterion, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Hong Kong, 2003.
[209] M.-H. Siu, G. Yu, H. Gish, An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, San Francisco, USA, pp. 189–192, 1992.
[210] F. Kubala, H. Jin, S. Matsoukas, L. Gnuyen, R. Schwartz, J. Machoul, The 1996 BBN byblos HUB-4 transcription system, Speech Recognition Workshop, pp. 90–93, 1997.
[211] P. Woodland, M. Gales, D. Pye, S. Young, The development of the 1996 HTK broadcast news transcription system, Speech Recorgnition Workshop, pp. 73–78, 1997.
[212] J. F. Lopez, D. P. W Ellis, Using acoustic condition clustering to improve acoustic change detection on broadcast news, Proc. International Conference on Speech and Language Processing, Beijing, China, 2000.
[213] D. Liu, F. Kubala, Fast speaker change detection for broadcast news transcription and indexing, Eurospeech-99, Vol. 3, Budapest, Hungary, pp. 1031–1034, 1999.
[214] S. Wegmann, F. Scattone,I. Carp,L. Gillick, R. Roth, J. Yamron, Dragon system's 1997 broadcast news transcription system, DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, USA, 1998.
[215] S. Tranter, D. Reynolds, Speaker diarization for broadcast news, ODYSSEY'04, Toledo, Spain, 2004.
[216] T. Hain, S. Johnson, A. Turek, P. Woodland and S. J. Young, S, Segment generation and clustering in the HTK broadcast news transcription system, DARPA Broadcast News Transcription and Understanding Workshop, pp. 133–137, 1998.
[217] L. Wilcox, F. Chen, D. Kimber, and V. Balasubramanian, Segmentation of speech using speaker identification, in IEEE Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP), vol. 1, pp. 161–164,1994.
[218] J. Ajmera, I. McCowan, and H. Bourlard, Speech/music segmentation using entropy and dynamism features in a HMM classification framework, Speech Commun., vol. 40, pp. 351–363, 2003.
[219] G. Lathoud and I. A. McCowan, Location based speaker segmentation, Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), vol. 1, pp. 176–179, 2003.
[220] M. Vescovi, M. Cettolo, R. Rizzi, A DP algoritm for speaker change detection, Eurospeech'03, 2003.
[221] J. Zdansky, J. Nouza, Detection of acoustic change-points in audio records via global BIC maximization and dynamic programming, Proc. International Conference on Speech and Language Processing, Lisbon, Portugal, 2005.
[222] M. Pwint, F. Sattar, A segmentation method for noisy speech using genetic algorithm, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, USA, 2005.
[223] G. Lathoud, I. McCowan and J. Odobez, Unsupervised location-based segmentation of multi-party speech, ICASSP-NIST Meeting Recognition Workshop, 2004.
[224] B. Fergani, M. Davy, A. Houacine, Speaker diarization using one-class support vector machines, Speech Communication 50, 2008.
[225] S. Furui, 50 years of progress in speech and speaker recognition, Proceedings of SPECOM 2005, Patras, Greece, 1–9, 2005.
[226] L. Rabiner and B. Huang. Fundamentals of Speech Recognition. Prentice Hall, 1993.
[227] J.-L. Gauvain, L. Lamel, Structuring Broadcast Audio for Information Access. EURASIP journal on Applied Signal Processing, pp. 140-150, 2003.

[228] StreamSage, Inc Phrase-Based Multimedia Information Extraction, Technical Report, Air Force Research Laboratory Information Directorate Rome Research Site Rome, New York , July 2006
[229] D. L. Hillard , Automatic Sentence Structure Annotation for Spoken Language Processing, Doctor of Philosophy,University of Washington, 2008.
[230] Boemie Deliverable 2.3, Semantics extraction from non-visual content tools: state-of-the-art report, 2007.
[231] K. Koskenniemi, Two-level morphology: a general computational model for word-form recognition and production. Publication No.11. Helsinki: University of Helsinki Department of General Linguistics. 1983.
[232] http://www.lingsoft.fi/doc/gertwol/
[233] W. Finkler, G. Neumann, Morphix: A Fast Realization of a Classification-Based Approach to Morphology, Proceedings of the 4th Austrian Artificial Intelligence Conference. 1988.
[234] D. Petitpierre, G Russell, MMORPH - The Multext Morphology Program. Multext deliverable report for the task 2.3.1, ISSCO, University of Geneva. 1995.
[235] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, M. Asahara, Japanese Morphological Analysis System ChaSen, version 2,0, Manual 2nd edition. 1999.
[236] http://www.xrce.xerox.com/competencies/content-analysis/fst/home.en.html
[237] http://www.issco.unige.ch//projects/MULTEXT.html
[238] E. Brill, A simple rule-based part of speech tagger. Proceedings of the Third Annual Conference on Applied Natural Language Processing, ACL. 1992.
[239] E. Brill, Unsupervised learning of disambiguation rules for part of speech tagging. Proceedings of the third ACL Workshop on Very Large Corpora. 1995.
[240] P. Tapanainen, A. Voutilainen, Tagging accurately: don't guess if you don't know. Technical Report, Xerox Corporation. 1994.
[241] D. Cutting, J. Kupiec, J. Pedersen, P. Sibun, A Practical Part-of-Speech Tagger, Proceedings of the 3rd conference on Applied Natural Language Processing (ANLP). 1992.
[242] H. Schmid, Probabilistic Part-of-Speech Tagging Using Decision Trees, International Conference on New Methods in Language Processing. Manchester. 1994.
[243] T. Brants, TnT - A Statistical Part-of-Speech Tagger, Proceedings of the 6th ANLP Conference, Seattle, WA. 2000.
[244] E. Brill, M. Marcus, Tagging an unfamiliar text with minimal human supervision. ARPA Technical Report. 1993.
[245] E. Brill, Transformation-based error-driven learning and natural language processing: A case study in part-of speech tagging. Computational Linguistics, 21(4), 543–566, 1995.
[246] S. Abney, Parsing by chunks. In Robert C. Berwick, Steven P. Abney, and Carol Tenny, editors, Principle-Based Parsing: Computation and Psycholinguistics, pages 257–278. Kluwer Academic Publishers, Boston, 1991.
[247] S. Abney, Partial parsing via finite-state cascade. Journal of Natural Language Engineering, 2(4): 337–344, 1996.
[248] L. Ramshaw, M. Marcus, Text chunking using transformation-based learning, David Yarovsky and Kenneth Church, editors, Proceedings of the Third Workshop on Very Large Corpora, pages 82–94, 1995.
[249] M. Osborne, Shallow parsing as part-of-speech tagging. In Claire Cardie, Walter Daelemans, Claire Nedellec, and Erik Tjong Kim Sang, editors, Proceedings of CoNLL-2000 and LLL-2000, pages 145–147, Lisbon, Portugal, 2000.
[250] M. Osborne, Shallow parsing using noisy and non-stationary training material. Journal of Machine Learning Research, 2(4):695–718, 2002.
[251] E. F. Tjong Kim Sang and S. Buchholz. Introduction to the conll-2000 shared task: Chunking in Claire Cardie, Walter Daelemans, Claire Nedellec, and Erik Tjong Kim Sang, editors, Proceedings of CoNLL-2000 and LLL-2000, pages 127–132. Lisbon, Portugal, 2000.
[252] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, US, 2002.

[253] M. Becker, W. Drozdzynski, H.U. Krieger, J. Piskorski, U. Schäfer, F. Xu., SProUT - Shallow Processing with Typed Feature Structures and Unification, Proceedings of ICON 2002 - International Conference on NLP, Mumbai, India, 2002.
[254] Y. Gao, B. Ramabhadran, J. Chen, H. Erdogan and M. Picheny, Innovative approach for large vocabulary name recognition, ICASSP 2001.
[255] C.-H. Tsai, N. Wang, P. Huang and J.-L. Shen, Open vocabulary Chinese name recognition with the help of character description and syllable spelling recognition, Proc. ICASSP-05, 2005.
[256] G. Stemmer, E. Nöth and H. Niemann, Acoustic modelling for foreign words in a German speech recognition system, Proc. Eurospeech 2001, 2001.
[257] B. Meison, S. Chen and P. Cohen, Pronunciation modeling for names of foreign origin, Proc. of ASRU-03, 2003.
[258] F. Beaufays, A. Sankar, S. Williams and M. Weintraub, (2003), Learning name pronunciations in automatic speech recognition systems, Proc. of 13th IEEE International Conference Tools with Artificial Intelligence, 2003.
[259] F. Kubala, R. Schwartz,R. Stone, R. Weischedel, Named entity extraction from speech. In DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, 1998.
[260] D. Miller, S. Boisen, R. Schwartz, R. Stone, R. Weischedel, Named entity extraction from broadcast news. In the sixth conference on Applied Natural Language Processing, 316–324, Seattle,WA, 2000.
[261] D. Palmer, M. Ostendorf, J. Burger Robust information extraction from automatically generated speech transcriptions. Speech Communication, 32, 95–109, 2000.
[262] L. Zhai, P. Fung ,R. Schwartz, M. Carpuat & D. Wu, Using nbest lists for named entity recognition from chinese speech. In the Proceedings of the HLT/NAACL 2004, Boston, MA , 2004.
[263] A. Maedche, G. Neumann and S. Staab, Bootstrapping an Ontology-Based Information Extraction System. In: Studies in Fuzziness and Soft Computing, Intelligent Exploration of the Web, Springer, 2002.
[264] H. Alani, S. Kim, D.E. Millard, M.J. Weal, W. Hall, P.H. Lewis, N.R. Shadbolt, Automatic Ontology-Based Knowledge Extraction from Web Documents. IEEE Intelligent Systems, 18(1), pp. 14-21, 2003.
[265] V. Lopez, E. Motta, Ontology-driven Question Answering in AquaLog, Proceedings of 9th international conference on applications of natural language to information systems (NLDB, 2004).
[266] H. M. Müller, E.E. Kenny, P.W. Sternberg, Textpresso: An ontology-based information retrieval and extraction system for biological literature, PLoS Biol 2, 2004.
[267] S. Nirenburg and V. Raskin, Ontological Semantics, MIT Press, 2004.
[268] M. Decker, M. Erdmann, D. Fensel, R. Studer, Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information, Database Semantics: Semantic Issues in Multimedia, pp. 351-369, 1999.
[269] C. Nédellec and A. Nazarenko, Ontologies and Information Extraction, 2005.
[270] D. Maynard, W. Peters and Y. Li, Metrics for Evaluation of Ontology-based Information Extraction, proceedings of EON2006, Evaluation of Ontologies for the Web, 4th International EON Workshop, 2006.
[271] C. Faloutsos, M.Flicke, W.Niblack, J.Ashley, Q.Huang, B.Dom, M.Gorkani, J.Hafine, D.Lee, D.Petkovic, D.Steele, P.Yanker, "Query by image and video content: The QBIC system", IEEE Computer 1995
[272] Y.Rui, T.S.Huang, S.Mehrotra, M.Ortega, "Automatic matching tool selection using relevance feedback in MARS", Proc. 2nd Int. Conf. Visual Inform. Syst., 1997.
[273] W.Niblack, R.Barber et al,. "The QBIC Project: Querying images by content using colour, texture and shape", Proc. SPIE Storage and Retrieval for Image and Video Databases, Feb 1994
[274] "Relevance feedback techniques in interactive content-based image retrieval systems", Proc. IEEE Workshop Content Based Access of Image and Video Libraries (in conjunction with IEEE CVPR'97) 1997
[275] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive contentbased image retrieval. IEEE Trans. Circuits Syst. Video Technol., 8(5):644–655, Sept. 1998. Special Issue on Segmentation, Description, and Retrieval of Video Content

[276] "Exploring video structures beyond the shots", Proc. IEEE Conf. Multimedia Computing and System. 1998
[277] J. Rocchio, "Relevance feedback in information retrieval", In The Smart System-Experiments In Automatic Document Processing, Englewood Cliffs, NJ: Prentice Hall, 313-323, 1971
[278] E. Ide, "New experiments in relevance feedback", In The Smart System-Experiments In Automatic Document Processing, Englewood Cliffs, NJ: Prentice Hall, 337-354, 1971
[279] G. Salton, "Relevance feedback and the optimization of retrieval effectiveness", In The Smart System-Experiments In Automatic Document Processing, Englewood Cliffs, NJ: Prentice Hall, 337-354, 1971
[280] Aggarwal, C.C., Wolf, J.L., Wu, K.L. & Yu, P.S. (1999). Horting Hatches an Egg: A New Graph-Theoretic Approach to Collaborative Filtering. In Proceedings of KDD-99 (pp. 201-212). San Diego, CA: ACM
[281] J. Rocchio, "Relevance feedback in information retrieval", In The Smart System-Experiments In Automatic Document Processing, Englewood Cliffs, NJ: Prentice Hall, 313-323, 1971
[282] X. S. Zhou and T. Huang. Relevance feedback in image retrieval: A comprehensive review. ACM Multimedia Systems Journal, Special Issue on CBIR, 8(6):536–544, 2003
[283] Y. Ishikawa, R. Subramanya, and C. Faloutsos. MindReader: Querying databases through multiple examples. In A. Gupta, O. Shmueli, and J. Widom, editors, Proc. of the 24th Int. Conf. on VLDB, pages 218–227, New York, NY,USA, Aug. 1998. Morgan Kaufmann Publishers
[284] K. Porkaew, K. Chakrabarti, and S. Mehrotra. Query refinement for multimedia similarity retrieval in MARS. In Proc. of the ACM Int. Conf. on Multimedia, pages 235–238, Orlando, Florida, 1999
[285] Y. Rui and T. S. Huang. Optimizing learning in image retrieval. In IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR-00), pages 236–245, Los Alamitos, June 2000. IEEE Computer Society Press
[286] G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, Tokio, 1983
[287] Cox I.J, Miller M.L, Minka T. P, Papathomas T.V, Yianilos P. N, "The Bayesian image retrieval system, PicHunter: theory, implementation and psychophysical experiments", IEEE Trans. On Image Processing, Vol. 9, No. 1, Jan 2000, pp. 20 – 37
[288] N. Vasconcelos and A. Lippman. Bayesian relevance feedback for content-based image retrieval. In IEEE Proc. of Workshop on Content-based Access of Image and Video Libraries, pages 63–67, 2000
[289] Wood, M., Campbell, N. and Thomas, B., "Iterative refinement by relevance feedback in content-based digital image retrieval," in ACM Multimedia'98, Bristol, UK, (1998)
[290] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In Proc. of the ACM Int. Conf. on Multimedia, pages 107–118. ACM Press, 2001
[291] Gloria Bordogna and Gabriella Pasi, "A user-adaptive neural network supporting a rule based relevance feedback", Fuzzy Sets and Systems, Vol. 82, No. 9, Spt 1996, pp. 201 – 211
[292] Tong Zhao, Lilian H.Tang, Horace H.S.Ip, Feihu Qi, "On relevance feedback and similarity measure for image retrieval with synergetic neural nets", Neurocomputing, Vol. 51, April 2003, pp. 105 – 124
[293] Douglass, R. J., "Description definition language (DDL), knowledge representation language for MPEG-7 DDL," ISO/IEC JTC1/SC29/WG11, Lancaster, UK, Feb. pp. 124 (1999). Ferman, A. M., Tekalp, A. M., Mehrotra, R., "Histogram-based color descriptors for multiple frame color characterization," ISO/IEC/JTC1/SC29/WG11, Lancaster, UK, Feb. pp. 529 (1999)
[294] M. Koskela, J. Laaksonen, E. Oja, "Comparison of Techniques for Content-Based Image Retrieval", Proceedings of SCIUA 2001, Bergen, Norway, June 2001
[295] K. Wu and K.-H. Yap, "Fuzzy relevance feedback in content-based image retrieval," Proc. Int. Conf. Information, and Signal Processing and Pacific-Rim Conf. Multimedia, Singapore, 2003.
[296] K. Wu and K.-H. Yap, "Fuzzy relevance feedback in content-based image retrieval," Proc. Int. Conf. Information, and Signal Processing and Pacific-Rim Conf. Multimedia, Singapore, 2003.
[297] Giorgio Giacinto and Fabio Roli, "Bayesian relevance feedback for content based image retrieval", Pattern Recognition, Vol. 37, No. 7, July 2004, pp. 1499 – 1508

[298] Zhong Su, Hongjiang Zhang, Li. S, Shaoping Ma, "Relevance Feedback in content based image retrieval: Bayesian framework, feature subspaces and progressive learning", IEEE Trans. on Image Processing, Vol. 12, No. 8, Aug 2003, pp. 924 – 937
[299] Zhong Su, Hongjiang Zhang, Li. S, Shaoping Ma, "Relevance Feedback in content based image retrieval: Bayesian framework, feature subspaces and progressive learning", IEEE Trans. on Image Processing, Vol. 12, No. 8, Aug 2003, pp. 924 – 937
[300] Chio-Ting Hsu, Chuech-Yu Li, "Relevance feedback using generalized Bayesian framework with region-based optimization learning", IEEE Trans. on Image Processing, Vol. 14, No. 10, Oct. 2005, pp. 1617 – 1631
[301] S. R. Gunn, "Support vector machines for classification and regression, technical report", Image Speech and Intelligent Systems Research Group , University of Southampton, 1997
[302] Y. Chen, X. S. Zhou, T. S. Huang, "One-class SVM for Learning in Image Retrieval", ICIP'2001, Thessaloniki, Greece, October 7-10, 2001
[303] Q. Tian, P. Hong, T. S. Huang, "Update relevant image weights for content-based image retrieval using support vector machines", IEEE International Conference on Multimedia and Expo, Hilton New York & Towers, New York, NY, July 30 - Aug. 2, 2000
[304] F. Jing, M. Li, Hong-Jiang Zhang, and B. Zhang "Relevance Feedback in Region-Based Image Retrieval", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 14, No. 5, May 2004
[305] F. Jing, M. Li, Hong-Jiang Zhang, B. Zhang ,"An Efficient and Effective Region-Based Image Retrieval Framework", IEEE Transactions on Image Processing , vol.13, no.5, May 2004
[306] F. Jing, M. Li, Hong-Jiang Zhang, and B. Zhang "Relevance Feedback in Region-Based Image Retrieval", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 14, No. 5, May 2004
[307] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases", IEEE International Conference on Computer Vision, pages 59-66, January 1998
[308] F. Jing, M. Li, Hong-Jiang Zhang, B. Zhang ,"An Efficient and Effective Region-Based Image Retrieval Framework", IEEE Transactions on Image Processing , vol.13, no.5, May 2004
[309] N. Davies, J. Landay, S. Hudson, and A. Schmidt. Guest Editors' Introduction: Rapid Prototyping for Ubiquitous Computing. IEEE Pervasive Computing, 4(4):15-17, 2005.
[310] R. Eberhart and J. Kennedy. A new optimizer using particle swarm theory. In 6th International Symposium on Micro Machine and Human Science, pages 39-43, 1995.
[311] M. Dorigo, V. Maniezzo, A. Colorni, F. Maffioli, G. Righini, and M. Trubian. Heuristics from nature for hard combinatorial optimization problems. International transactions on operational research, 3(1):1-21, 1996.
[312] R. Eberhart and J. Kennedy. A new optimizer using particle swarm theory. In 6th International Symposium on Micro Machine and Human Science, pages 39-43, 1995.
[313] D. Dasgupta. Artificial Immune Systems and their Applications. Springer-Verlag, Heidelberg, Germany, 1999.
[314] D. Fogel. An introduction to simulated evolutionary optimization. IEEE Trans. on Neural Network, 5(1):3-14, January 1994.
[315] A. Grigorova, F. B. De Natale, C. Dagli, T. S. Huang, "Content-Based Image Retrieval by Feature Adaptation and Relevance Feedback", IEEE Trans. on Multimedia Vol. 9, No. 6, Oct. 2007, pg. 1183 - 1192
[316] A. Grigorova, F. B. De Natale, C. Dagli, T. S. Huang, "Content-Based Image Retrieval by Feature Adaptation and Relevance Feedback", IEEE Trans. on Multimedia Vol. 9, No. 6, Oct. 2007, pg. 1183 - 1192
[317] A.Dong, B. Bhanu, "Active Concept Learning in Image Databases", IEEE Tran. Systems, Man and Cybernetics, Vol. 35, No. 3, June 2005, pg. 450 - 466.
[318] P. Yin, B. Bhanu, K. Chang, and A. Dong, "Improving retrieval performance by long-term relevance information," in Proc. Int.Conf. Pattern Recognition, vol. III, Aug. 2002, pp. 533–536.
[319] J. Fournier and M. Cord, "Long-term similarity learning in contentbased image retrieval," in Proc. IEEE Int. Conf. Image Processing, vol. 1, Sep. 2002, pp. 441–444.
[320] B. Bhanu and A. Dong, "Concept learning with fuzzy clustering and relevance feedback," Eng. Applicat. Artif. Intell., vol. 15, pp. 123–138, Apr. 2002.

[321] N. Vasconcelos, "Bayesian Models for Visual Information Retrieval," Ph.D. dissertation, MIT, Cambridge, 2000.
[322] On the complexity of probabilistic image retrieval," in Proc. IEEE Int. Conf. Computer Vision, vol. 2, Jul. 2001, pp. 400–407.
[323] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in Proc. IEEE Int. Conf. Computer Vision, vol. 2, 2001, pp. 408–415. G. McLachlan and D. Peel, Finite Mixture Models. New York: Wiley, 2000.
[324] P. Muneesawang, L. Guan, "An Interactive Approach for CBIR Using a Network of Radian Basis Function", IEEE. Trans. on Multimedia, Vol. 6, No. 5, Oct. 2004
[325] S. Haykin, Neural Networks: A Comprehensive Foundation. New York: Macmillan, 1994.
[326] J. Moody and C. J. Darken, "Fast learning in networks of locally-tuned processing units," Neural Computat., vol. 1, no. 2, pp. 281–294, 1989.
[327] F. Jing, M. J. Li, H-J. Zhang, "An Efficient and Effective Region-Based Image Retrieval Framework", IEEE. Trans. On Image Processing, Vol. 13, No. 5, May 2004.
[328] T. P. Minka and R. W. Picard, "Interactive learning using a society of models," Pattern Recognit., vol. 30, no. 4, pp. 565–581, Apr. 1997.
[329] M. E. Wood, N. W. Campbell, and B. T. Thomas, "Iterative refinement by relevance feedback in content based digital image retrieval," in Proc. 5th ACM International Multimedia Conference (ACM Multimedia 98), Bristol, U.K., Sept. 1998, pp. 13–20.
[330] D. Djordjevic, E. Izquierdo, "An Object - and User- Driven System for Semantic Based Image Annotation and Retrieval", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 17, No. 3, March 2007.
[331] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," Proc. Comput. Vis. Pattern Recognit., pp. 264–271, 2003.
[332] S. Agarwal and D. Roth, "Learning a sparse representation for object detection," in Proc. Eur. Conf. Comput. Vis., 2002, pp. 113–130.
[333] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of keypoints," in Proc. Eur. Conf. Comput. Vis., Int. Workshop Statistical Learning in Computer Vision, 2004, pp. 59–74.
[334] A. Gounaris et al, "Knowledge Assisted Multimedia Analysis", Technical Report, 2007.
[335] W. Al-Khatib, Y. F. Day, A. Ghafoor, P. B. Berra, "Semantic Annotation of Images and Videos for Multimedia Analysis", 2nd European Semantic Web Conference (ESWC), Herakleion, Greece, 2005
[336] R. Tansley, C. Bird, W. Hall, P. Lewis, M. Weal. "Automating the linking of content and concept". In Proc. ACM Int. Multimedia Conf. and Exhibition (ACM MM-2000), 2000.
[337] I. Kompatsiaris, V. Mezaris and M.G. Strintzis. „Multimedia content indexing and retrieval using an object ontology". Multimedia Content and Semantic Web-Methods, Standards and Tools, Wiley, NY, 2004.
[338] N. Maillot, M. Thonnat and C. Hudclot. "Ontology based object learning and recognition: Application toimage retrieval". Proc. of IEEE Internaional Conference on Tools with Artificial Intelligence (ICTAI), pp. 620-625, Boca Raton, FL, USA, 2004.
[339] J. Assfalg, M. Berlini, A. Del Bimbo, W. Nunziat, P. Pala. Soccer Highlights Detection and Recognition using HMMs. IEEE International Conference on Multimedia & Expo (ICME), 825-828, 2005
[340] J. Zhao, Y. Shimazu, K. Ohta, R. Hayasaka, Y. Matsushita. An Outstandingness Oriented Image Segmentation and its Applications. In Proc. of the International Symposium on Signal Processing and its Applications, 1996
[341] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V.K. Papastathis, M.G. Strintzis. Knowledge-Assisted Semantic Video Object Detection. IEEE Transactions, CSVT, Special Issue on Analysis and Understanding for Video Adaptation, 15(10), 1210–1224, 2005
[342] L. Hollink, S. Little, J. Hunter. Evaluating the Application of Semantic Inferencing Rules to Image Annotation. 3rd International Conference on Knowledge Capture (K-CAP05), Banff, Canada, 2005
[343] Y. Wang, F. Makedon, J. Ford, L. Shen, D. Golding. Generating Fuzzy Semantic Metadata Describing Spatial Relations from Images using the R-Histogram. JCDL '04, June 7-11, Tucson, Arizona, USA, 2004.

[344] L. Hollink, G. Nguyen, G. Schreiber, J. Wielemaker, B. Wielinga, M. Worring. Adding Spatial Semantics to Image Annotations. In Proc. of International Workshop on Knowledge Markup and Semantic Annotation, ISWC, 2004
[345] S. Skiadopoulos, C. Giannoukos, N. Sarkas, P. Vassiliadis, T. Sellis, M. Koubarakis. 2D topological and direction relations in the world of minimum bounding circles. IEEE Transactions on Knowledge and Data Engineering, 17(12), 1610-1623, 2005
[346] B. Edmonds. The Pragmatic Roots of Context. In Proc. of the 2nd International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-99), LNAI, vol. 1688, pp. 119-132, Berlin, Springer, 1999
[347] Th. Athanasiadis, V. Tzouvaras, K. Petridis, F. Precioso, Y. Avrithis, I. Kompatsiaris. Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content. In Proc. of SemAnnot '05, Galway, Ireland, 2005
[348] J. Zhao, Y. Shimazu, K. Ohta, R. Hayasaka, Y. Matsushita. An Outstandingness Oriented Image Segmentation and its Applications. In Proc. of the International Symposium on Signal Processing and its Applications, 1996
[349] C. Hudelot and M. Thonnat, "A Cognitive Vision Platform for Automatic Recognition of Natural Complex Objects", in Proc. Of 15th IEEE International Conference on tools for Artificial Intelligence (ICTAI), 2003.
[350] J. Hunter, J. Drennan, and S. Little, "Realizing the hydrogen economy through semantic web technologies", IEEE Intelligent Systems, vol. 19, no. 1, pp. 40-47, 2004.
[351] S. Little and J. Hunter, "Rules-by-Example – a novel approach to semantic indexing and querying of images", in International Semantic Web Conference, pp. 534-548, 2004
[352] J.-P. Schober, T. Hermes, and O. Herzog, "Content-based iamge retrieval by ontology-based object recognition", in Proc. KI-2004 Workshop on Applications of Description Logics, (ADL-2004), V. Haarslev, C. Lutz, and R. Moller, Eds., 2004.
[353] B. Neumann and R. Möller, "On scene interpretation with description logics", Tech. Rep. FBI-B-257/04,University of Hamburg, Computer Science Department, 2004
[354] R. Möller, B. Neumann, and M. Wessel, "Towards Computer Vision with Description Logics: Some Recent Progress",. In Proceedings Integration of Speech and Image Understanding, Corfu, Greece, pages 101-115, 1999
[355] C. Meghini, F. Sebastiani, and U. Straccia, "A model of multimedia information retrieval",J. ACM, vol. 48, no. 5, pp. 909-970, 2001
[356] L.Hollink, M. Worring and A.Th. Schreiber. Building a Visual Ontology for Video Retrieval. Accepted as a short paper in ACM MultiMedia 2005
[357] Anthony Hoogs, Jens Rittscher, Gees Stein, John Schmiederer: Video Content Annotation Using Visual Analysis and a Large Semantic Knowledgebase. CVPR (2) 2003: 327-334.
[358] A.Dorado, E.Izquierdo, "Exploiting problem domain knowledge for accurate building image classification", 3rd International Conference on Image and Video Retrieval 2004, Vol. 3115, July 2004, pp. 199 – 206
[359] M. Petkovic and W. Jonker, W., "Content-Based Video Retrieval by Integrating Spatio-Temporal and Stochastic Recognition of Events", In Kluwer Academic Publishers, Boston, Hardbound, ISBN 1-4020-7617-7, 168 pp, 2003
[360] X. Zhu, X. Wu, A. K. Elmagarmid, Z. Feng, and L. Wu, "Video data mining: Semantic indexing and event detection from the association perspective," IEEE Trans. Knowl. Data Eng., vol. 17, no. 5, pp. 665–677, May 2005.
[361] R. Leonardi, P. Migliorati, and M. Prandini, "Semantic indexing of soccer audio-visual sequences: A multimodal approach based on controlled Markov chains," IEEE Trans. Circuits Syst. Video Technol., vol.14, no. 5, pp. 634–643, May 2004.
[362] S. Dagtas and M. Abdel-Mottaleb, "Extraction of TV highlights using multimedia features," in Proc. IEEE Int. Workshop on Multimedia Signal Processing, Cannes, France, 2001, pp. 91–96.
[363] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," IEEE Trans. Image Process., vol. 12, no. 7, pp. 796–807, Jul. 2003.
[364] M. Han, W. Hua, W. Xu, and Y. Gong, "An integrated baseball digest system using maximum entropy method," in Proc. ACMInt. Conf. Multimedia, Juan les Pins, France, 2002, pp. 347–350.

[365] Z. Xiong, X. Zhou, Q. Tian, Y. Rui, and T. S. Huang, "Semantic retrieval of video," IEEE Signal Process. Mag. (Special Issue on Semantic Retrieval of Multimedia), vol. 23, no. 2, pp. 18–27, Mar. 2006.
[366] X. Wu, C.-W. Ngo, and Q. Li, "Threading and auto-documentary in news videos," IEEE Signal Process. Mag., vol. 23, no. 2, pp. 59–68, Mar. 2006.
[367] M. R. Naphade and J. R. Smith, "On the detection of semantic concepts at TRECVID," in Proc. 12th ACM Int. Conf. Multimedia, New York, 2004, pp. 660–667.
[368] V. S. Tseng, C.-J. Lee, and J.-H. Su, "Classify by representative or associations (CBROA): A hybrid approach for image classification," in Proc. 6th Int. Workshop on Multimedia Data Mining: Mining Integrated Media and Complex Data, Chicago, IL, Aug. 2005, pp. 37–53
[369] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pp. 696–710, Jul. 1997.
[370] D. Sadlier and N. E. O'Connor, "Event detection in field-sports video using audio-visual features and a support vector machine," IEEE Trans. Circuits Syst. Video Technology, vol. 15, no. 10, pp. 1225–1233, Oct. 2005.
[371] A. Amir et al., "IBM research TRECVID-2003 video retrieval system," in NIST TRECVID, 2003.
[372] C. G. M. Snoek, M. Worring, J. C. Gemert, J.-M. Geusebroek, and A. W. M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in Proc. ACM Int. Conf. Multimedia, Santa Barbara, CA, 2006, pp. 421–430
[373] B. Han, Support Vector Machines Center for Information Science and Technology, Temple University, Philadelphia, PA, 2003 [Online]. Available: http://www.ist.temple.edu/~vucetic/cis526fall2003/lecture8.doc
[374] J. R. Quinlan, C4.5: Programs for Machine Learning. San Francisco, CA: Morgan Kaufmann, 1993.
[375] M-L. Shyu, Z. Xie, M. Chen, S-C Chen, "Video Semantic Event/Concept Detection Using a Subspace - Based Multimedia Data Mining Framework", IEEE Trans. On Multimedia, Vol. 10, NO. 2, Feb. 2008
[376] M. Chen, S.-C. Chen, M.-L. Shyu, and K. Wickramaratna, "Semantic event detection via temporal analysis and multimodal data mining," IEEE Signal Processing Mag. (Special Issue on Semantic Retrieval of Multimedia), vol. 23, no. 2, pp. 38–46, Mar. 2006.
[377] Y. Wang, Z. Liu, J-C. Huang, "Multimedia Content Analysis using both Audio and Visual Clues", IEEE Signal Processing Magazine, pg. 12 - 36, Nov. 2000
[378] S. Luke, L. Spector, D. Rager, and J. Hendler, "Ontology-based Web Agents", in proc. of the First International Conference on Autonomous Agents (Agents97). W. L. Johnson, ed. Association for Computing Machinery, New York, 1997, pp. 59–66.
[379] S. Decker, M. Erdmann, D. Fensel, and R. Studer, "Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information", in R. Meersman et al. (eds.), Semantic Issues in Multimedia Systems, Kluwer Academic Publisher, Boston, 1999.
[380] S. Handschuh, S. Staab, and A. Maedche, "CREAM – Creating Relational Metadata with a Component-Based, Ontology-Driven Framework", in proc. of the 1st Int. Conf. on Knowledge Capture (K-CAP 2001), Victoria B.C., Canada, pp. 76–83, October 2001.
[381] V. R. Benjamins, D. Fensel, S. Decker, and A. Gómez-Pérez, "(KA)2: Building Ontologies for the Internet: a Mid Term Report", International Journal of Human-Computer Studies (IJHCS), Vol. 51, pp. 687–712, 1999.
[382] P. Kogut and W. Holmes, "AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages", in Proc. of the 1st Int. Conf. on Knowledge Capture (K-CAP 2001), Workshop on Knowledge Markup and Semantic Annotation, Victoria B.C., Canada, 2001.
[383] Y. Li, L. Zhang, and Y. Yu, "Learning to Generate Semantic Annotation for Domain Specific Sentences", in knowledge markup and semantic annotation workshop K-CAP 2001, 2001.
[384] J. Kahan, M.-R. Koivunen, E. Prud'Hommeaux, and R. R. Swick, "Annotea: An Open RDF Infrastructure for Shared Web Annotations", in proc. of the 10th International World Wide Web Conference (2001), Hong Kong, ACM Press, May 1-5, 2001, pp. 623–632.
[385] C. Goble, S. Bechhofer, L. Carr, D. De Roure, and W. Hall, "Conceptual Open Hypermedia = The Semantic Web?", in proc. of the 2nd Int. Workshop on the Semantic Web (SemWeb2001), Hong Kong, May 2001.

[386] R. Baumgartner, S. Flesca, and G. Gottlob, "Visual Web Information Extraction with Lixto", in proc. of 27th conference on Very Large Data Bases (VLDB 2001), Rome, Italy, 11-14 September 2001, pp. 119–128

[387] F. Ciravegna, S. Chapman, A. Dingli, and Y. Wilks, "Learning to Harvest Information for the Semantic Web", in proc. of the 1st European Semantic Web Symposium (ESWS 2004), Heraklion, Greece, May 10-12, 2004.

[388] A. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut, and Y. Warke, "Managing Semantic Content for the Web", IEEE Internet Computing, Vol. 6(4), pp. 80–87, 2002.

[389] J. Domingue, M. Dzbor, and E. Motta, "Semantic Layering with Magpie",. in: Staab, S. & Studer, R. (eds.), Handbook on Ontologies in Information Systems, Springer, Verlag, 2003.

[390] A. Vasilakopoulos, M. Bersani, and W. J. Black, "A Suite of Tools for Marking Up Textual Data for Temporal Text Mining Scenarios", in proc. of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, 24-30 May 2004.

[391] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, K. S. McCurley, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zienberer, A Case for Automated Large Scale Semantic Annotation. Journal of Web Semantics, Vol. 1 (1), December 2003.

[392] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov, "KIM – Semantic Annotation Platform", in poc. of 2nd International Semantic Web Conf. (ISWC 2003), 20-23 October 2003, Florida, USA, Springer-Verlag Berlin Heidelberg, Vol. 2870, pp. 834–849, 2003.

[393] A. Wessman, S. W. Liddle, and D. W. Embley, "A generalized framework for an ontology-based data-extraction system", in proc. of 4th Int. Conf. on Information Systems Technology and its Applications, pp. 239–253, 2005.

[394] D. W. Embley, D. M. Campbell, Y. S. Jiang, S. W. Liddle, D. W. Lonsdale, Y.-K. Ng, and R. D. Smith, "Conceptual-model-based data extraction from multiple-record Web pages", Data & Knowledge Engineering, Vol. 31 (3), pp. 227–251, November 1999.

[395] O. Valkeapää and E. Hyvönen, "A Browser-based Tool for Collaborative Distributed Annotation for the Semantic Web", in proc. of 5th International Semantic Web Conference, Semantic Authoring and Annotation Workshop, November 2006.

[396] Y. Bodain and J.-M. Robert, "Developing a robust authoring annotation system for the Semantic Web", in proc. of the 7th IEEE International Conference on Advanced Learning Technologies (ICALT 2007), pp. 391–395, 2007.

[397] I. Muslea, S. Minton, and C. A. Knoblock, "Active learning with strong and weak views: A case study on wrapper induction", in proc. 18th Int. Joint Conference on Artificial Intelligence, pp. 415–420 2003.

[398] D. Freitag and N. Kushmerick, "Boosted wrapper induction", in proc. of 17th National Conference on Artificial Intelligence, pp. 577–583, 2000.

[399] N. Kiyavitskaya, N. Zeni, L. Mich, J. R. Cordy, and J. Mylopoulos, "Text Mining through Semi-Automatic Semantic Annotation", in proc. of the 6th International Conference on Practical Aspects of Knowledge Management (PAKM 2006), LNCS, Springer-Verlag, Vol. 4333, pp. 143–154, 2006.

[400] K. Kerremans, Towards multilingual, termontological support in ontology engineering, In Proceedings of the Workshop on Terminology, Ontology and Knowledge Representation, Lyon, France, 2004

[401] W. Peters, E. Montiel-Ponsoda, G. A Aguado de Sea, Localizing Ontologies in OWL, OntoLex 2007

[402] I. Terziev, A. Kiryakov, D. Manov, Base Upper-level Ontology (BULO) Guidance Deliverable 1.8.1, SEKT project, July 2005, http://proton.semanticweb.org/D1_8_1.pdf

[403] Vossen, P. (eds) 1998 EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht

[404] Vossen, P., N. Calzolari, G. Adriaens, A. Sanfilippo, Y. Wilks (eds.) 1997, Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Proceedings of the ACL/EACL-97 workshop, Madrid, July 12th, 1997.