Journal of Visual Languages and Computing 32 (2016) 35–41

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

Semantic processing of multimedia data for e-government applications

Flora Amato [a,*], Francesco Colace [b], Luca Greco [b], Vincenzo Moscato [a], Antonio Picariello [a]

[a] Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione. University of Naples Federico II, Italy
[b] Dipartimento di Ingegneria dell'Informazione, Ingegneria Elettrica e Matematica Applicata. University of Salerno, Italy

ARTICLE INFO

Available online 29 November 2015

Keywords:
Ontology learning
Natural languages processing
Information systems
Multimedia

ABSTRACT

Knowledge management has become a challenge for almost all e-government applications where the efficient processing of large amounts of data is still a critical issue. In the last years, semantic techniques have been introduced to improve the full automatic digitalization process of documents, in order to facilitate the access to the information embedded in very large document repositories. In this paper, we present a novel model for multimedia digital documents aiming at improve effectiveness of digitalization activities within an information system supporting e-government organizations. At the best of our knowledge, the proposed model is one of the first attempts to give a single and unified characterization of multimedia documents managed by e-government applications, whereas semantic procedures and multimedia facilities are used for the transformation of unstructured documents into structured information. Furthermore, we define an architecture for the management of multimedia documents "life cycle", which provides advanced functionalities for information extraction, semantic retrieval, indexing, storage, presentation, together with long-term preservation. Preliminary experiments concerning an e-health scenario are finally presented and discussed.

1. Introduction

E-government (e-gov) activities are devoted to improve efficiency, expensiveness and accessibility of public administration services: *digitalization* is surely one of the most important tasks within this kind of context.

Indeed, the core aspect related to an effective digitalization process is the idea standing beyond the common document concept: it can be defined as "the representation of acts, facts and figures directly made or by means of electronic processing, and stored into an intelligible support".[1] In other words, a document can be seen as a set of multimedia objects (e.g. text and images) that, according to their relative positions within the support, determines the shape and, consequently the structure of the document.

In addition, during the various processing phases, depending on the particular application domain, a document is computed and eventually stored into different kinds of media.

In order to properly manage documents, *Document Management Systems* (DMS) have been introduced. Initially, they were used for converting paper documents into electronic images. Nowadays, DMS are becoming the basis of the majority of information systems, giving the users an access to "company knowledge", providing efficient and effective retrieval, reducing error rates in documents manipulation and thus improving overall business performances.

With the advent of *Semantic Web* and the adoption of standards for knowledge representation, DMS evolved from simple search engines, towards more complex systems able to integrate semantic technologies for information extraction and retrieval. Such systems, however, are

* Dep. of Ingegneria Elettrica e delle Tecnologie dell'Informazione. University of Naples Federico II. Via Claudio, 21. Napoli 80125. Italy. Tel.: +39 081 7683851.
E-mail addresses: flora.amato@unina.it (F. Amato), fcolace@unisa.it (F. Colace), lgreco@unisa.it (L. Greco), vmoscato@unina.it (V. Moscato), picus@unina.it (A. Picariello).

[1] This definition accords with the Italian civil law [1].

http://dx.doi.org/10.1016/j.jvlc.2015.10.012
1045-926X/© 2015 Elsevier Ltd. All rights reserved.

limited to provide additional semantic functionalities to existent document management features.

In the literature, there are a variety of semantic-based approaches to model multimedia content focusing on single type of media, but there exist only few proposals [2] for processing more complex multimedia documents as required by e-gov applications. Generally speaking, the main goal of a semantic-based processing is to structure input documents and to allow automatic retrieval of targeted information on the base of a formal representation of the related domain.

In this paper, we propose a new model of multimedia documents that meet with the specific requirements of e-gov applications, allowing a semantic processing of the related digital contents that can be exploited from various perspectives such as presentation, indexing, integration, storage, retrieval and so on. In particular, our model allows: (i) documents structuring; (ii) automatic information extraction from digital documents; (iii) semantic retrieval; (iv) semantic interpretation of the relevant information presented in the document; (v) storing; and (vi) long term preservation.

From a technical perspective, the proposed system combines Object-Relational Database (ORDBMS) technologies, Natural Language Processing (NLP) techniques, proper domain and structural ontologies, and inference rules in order to automatically extract significant concepts from each document (document annotation) and to provide semantic querying facilities [2] (retrieval is improved by “enriching and then refining” the set of the retrieved documents by using reasoning techniques on the ontological relations).

We consider the *e-health* domain as a suitable case study: it implies a massive document processing that must be performed in reliable, effective and error-free way. In particular, we focus our attention on the semantic processing of *Electronic Clinical Records*[2] that, as well known, can contain several types of multimedia contents.

The paper is organized as in the following. Section 2 reports a brief review describing the main DMS and multimedia information management technologies. Section 3 describes the proposed framework for manage multimedia documents. Section 4 outlines some implementation details for our system, while Section 5 presents some preliminary experimental results for e-health domain. Finally, Section 6 discusses conclusions and future work.

2. Related works

Starting from the 1980s, a number of vendors began to develop systems to manage paper-based documents. More recently, Document Management Systems (DMS) have been dedicated to the management of digital documents, providing a set facilities for document processing as storage, versioning, metadata management, security, as well as indexing and retrieval capabilities. Nowadays, DMS are evolving to integrate semantic functionalities, including advanced features for contents management like semantic search (as EUNOMOS, a knowledge management system for legal documents).

In the last years, numerous projects for document management in several specialist domains are presented, as the ASTREA Project realized by the Judicial Systems Research Institute (IRSIG), the TAPA Project realized for the Anti-trust Authority (AGCM) and the ESTRELLA Project (European project for Standardized Transparent Representations in order to Extend Legal Accessibility) financed by the European Union.

They combine several NLP and machine learning techniques to extract structured information form data (text) and Semantic Web technologies to support semantic retrieval.

Concerning, the state of the art in multimedia information management system, one of the main research objectives is the automatic indexing of multimedia data on the basis of their content in order to make query processing easier, more effective and efficient.

In particular, the major challenges in developing reliable image database systems lie in the capability of such systems in extracting relevant information on the base of image visual content and semantics expressed by means of simple attributes (metadata), tags or keywords.

Traditionally, the problem of finding relevant images to the users on the base of visual content is solved using low-level image global descriptors (color, texture and shape features) for which automatic extraction methods are available, see [5] for details. More recently, it has been realized that such global descriptors are not suitable to describe the actual objects within the images and their associated semantics. Two main approaches have been proposed to cope with this deficiency: the first approach segments the image into multiple regions, and different descriptors are built for each region [5]; the second approach exploits salient points identification techniques [6]. Finally, more recent systems [9] have as goal the automatic classification of images on the base of low-level features and high-level human annotations.

3. The proposed framework

3.1. Document model

In order to manage the different kinds of multimedia data, their relations and the particular structure imposed by e-gov applications, the adopted document model uses three different representation layers, as described in the following.

Data management layer: describes the semantic content of each single multimedia objects composing the document (such as a text fragment or an image), providing functionalities for managing each single media; as an example, information extraction and indexing over text and images are performed in this layer.

[2] According to the *International Organization for Standardization* (ISO) definition, an electronic clinical record means a repository of patient data in digital form, stored and exchanged securely, and accessible by multiple authorized users. It contains retrospective, concurrent, and prospective information, and its primary purpose is to set objectives and planning patient care, document the delivery of care and assess the outcomes of care.