

COMPUTERS ON MEDICAL DIAGNOSIS

MONICA LUPȘOR, RADU BADEA, CRISTIAN VICAȘ, SERGIU NEDEVSCHI, HORIA ȘTEFĂNESCU,
MIHAI SOCACIU, CORINA RADU , DANA CRIȘAN, MIRCEA GRIGORESCU
The performance of the attenuation coefficient computed on the ultrasound image in
quantifying liver steatosis in diffuse liver disease patients. 199

CRISTIAN VICAȘ, SERGIU NEDEVSCHI, MONICA LUPȘOR, RADU BADEA
Fibrosis detection using ultrasound and serological markers as features for additive
logistic regression. 205

MIHAI SOCACIU, RADU BADEA, CLAUDIA HAGIU, SILVIU IOBAGIU, ADELA GOLEA, LIDIA
CIOBANU, PAULINA MITREA, DELIA MITREA
Ultrasound Assessment Techniques and Quantification of Bowel Inflammation in
Several Intestinal Diseases: Initial Experience. 211

DELIA MITREA, SERGIU NEDEVSCHI, MONICA LUPȘOR, MIHAI SOCACIU, RADU BADEA
Improving the imagistic textural model of the hepatocellular carcinoma through
dimensionality reduction techniques and classifier combinations. 217

ANCA CIURTE AND SERGIU NEDEVSCHI
Abdominal Lesion Segmentation within CT images. A survey. 223

MARIA CRIȘAN, CARLO CATTANI, RADU BADEA, RODICA COSGAREA, SORIN DUDEA, MONICA
LUPȘOR, ADRIAN STIR, PAULINA MITREA, DELIA MITREA
Complex Histological, Ultrasonography and Clinical Studies in early noninvasive
diagnosis of the Photo-induced Cutaneous Senescence and in the Photo-induced
Skin Cancers, using Computerized Imaging, Modern Biotechnology and
Mathematical Modelling Methods. 230

ALEXANDRU I. MITREA, OCTAVIAN MIRCEA GURZĂU
Discretization Methods for obtaining Energy-Minimizing Deformable Models. 234

GHEORGHE SEBEȘTYEN, LORAND KRUCZ, GAVRIL SĂPLĂCAN, MIRCEA RUSU
Computer-aided Remote Patient Monitoring with Personal Healthcare Units. 239

LUCIA VICAȘ, DIANA PELEA, CRISTIAN VICAȘ
Software solution to quality assessment in medical analytical laboratories. 243

DIGITAL LIBRARIES, E-CONTENT MANGEMENT AND E-LEARNING

DOINA BANCIU
“e-Romania” from Concept to Implementation 246

DORA COARDOȘ, VASILE COARDOȘ, CORNELIA LEPĂDATU
Integrated On-line System for Management of the National Retrospective
Bibliography – SIMBNR 249

ANA-MARIA BOROZAN, ILEANA TRANDAFIR
Ontology-based content and learning personalization in an e-learning system. 254

DRAGOS BARBU, DRAGOS NICOLAU, LAURA CIOCOIU, IONUT PETRE Tools for Digital Content Management.	258
GHEORGHE SEBEȘTYÉN-PÁL, BOGDAN MOSCALIUC, CHRISTIAN LUDOVIC NAGY Revalorization of National Patrimony through Digital Content Management	261
RODICA POTOLEA, CAMELIA VIDRIGHIN BRATU, FLORIN TRIF Intelligent Component for adaptive E-learning Systems.	267
CLAUDIU VASILE KIFOR, LUCIAN LOBONȚ, CONSTANTIN OPREAN Managing Knowledge and Quality in Higher Education Study Programmes.	273
IOAN SALOMIE, CRISTINA POP, TUDOR CIOARĂ, IONUȚ ANGHEL, MARIUS GRINDEAN, IOAN DRĂGAN, LIVIA ARDELEAN, ION CARJA Applying Rule-based Knowledge Acquisition on Historical Documents.	277
MIHAELA DINȘOREANU, TUDOR CIOARĂ, IONUȚ ANGHEL, LAURA DINȘOREAN, MARIUS GEORGIU, ION CARJA, IOAN DRĂGAN, LIVIA ARDELEAN Information Extraction from Romanian Archival Documents.	283
MIHAELA DINȘOREANU, IOAN SALOMIE, CORNEL RAȚ, SORIN SUCIU, TUDOR CIOARĂ, IONUȚ ANGHEL, IOAN DRĂGAN, LIVIA ARDELEAN, ION CĂRJA Hierarchical Data Models for the ArhiNet Ontology Representation and Processing. . .	289
IOAN SALOMIE, CORNEL RAȚ, SORIN SUCIU, CRISTINA POP, VIORICA CHIFU, IOAN DRĂGAN, LIVIA ARDELEAN, ION CĂRJA Distributed, Agent-based, Semantic Reasoning Engine for Ontologies A Case-study in the Archival Domain.	295

The performance of the attenuation coefficient computed on the ultrasound image in quantifying liver steatosis in diffuse liver disease patients

Monica Lupsor, Radu Badea,
Horia Ștefănescu, Mihai Socaciu, Corina Radu,
Dana Crisan, Mircea Grigorescu
3rd Medical Clinic, University of Medicine and
Pharmacy, Cluj Napoca, Romania,
monica.lupsor@umfcluj.ro ; rbadea@umfcluj.ro

Cristian Vicas, Sergiu Nedevschi

Technical University Cluj Napoca, Romania
{cristian.vicas;sergiu.nedevschi}@cs.utcluj.ro

Abstract – The current study aims to evaluate the performance of attenuation coefficient (AC) in quantifying liver steatosis in diffuse liver diseases, compared to the classical ultrasonographic (US) examination. On the whole, we found a significant difference of the AC mean according to the severity of steatosis: -0.0008 (in patients without steatosis, S0), -0.0538 (S1 - steatosis 6-33%), -0.1184 (S2 - steatosis 34-66%) and -0.1484 respectively (S3 - steatosis >66%). The cutoff values of AC predictive for each grade of steatosis were: ≤ -0.0471 for the prediction of $S \geq 1$ (Se 50.73%, Sp 92.72%, PPV 88.8%, NPV 62.4%), ≤ -0.0821 for the prediction of $S \geq 2$ (Se 79.66%, Sp 76.47%, PPV 43.1%, NPV 94.4%) and ≤ -0.1474 for the prediction of $S \geq 3$ (Se 66.67%, Sp 91.60%, PPV 32.0%, NPV 97.9%). AC could be used to develop an imaging method for the detection of steatosis grade, that is less operator dependent, even in the presence of other histological alterations. This method could make the classical US exam more objective, it could increase its specificity and it could be useful to a proper diagnosis and evolution estimation in diffuse liver diseases.

Liver steatosis, attenuation coefficient, noninvasive diagnosis

I. INTRODUCTION

The diagnosis of liver steatosis has for a long time been based on the morphologic examination of liver tissue obtained by needle liver biopsy (LB). Unfortunately, this diagnostic method is an invasive one, having potential secondary effects and a certain intra and interobserver variability. Furthermore, LB does not always accurately reflect the real intrahepatocyte fat content when its distribution within the liver is inhomogeneous [1, 2]. Under these circumstances, the importance of finding non-invasive diagnostic methods has increased progressively. Among these, the imaging techniques (ultrasonography, computer tomography, magnetic resonance) have, besides being non-invasive, the advantage of assessing the entire organ. Ultrasonography (US) is the most used method, due to its wide spread availability, low cost and lack of secondary effects [3].

US changes found in steatosis are apparent only at more than 15-20% fat content in the hepatocytes; these are hepatomegaly, increase of liver echogenicity ("bright" liver), attenuation of ultrasounds in subcapsular areas, difficult visualising of the portal vein walls, the gall bladder and liver capsule, the apparent dilation of vessels (especially suprahepatic ones) and a false transonic aspect of the right kidney in comparison to the liver parenchyma [4, 5, 6].

US plays an important part in the diagnosis of steatosis, but it cannot assess accurately the degree of fat

infiltration nor can it totally differentiate it from fibrosis. The performance of the method varies considerably from one study to the next [7-9], possibly because of other concomitant histological alterations (inflammation, fibrosis) that can alter the ultrasound image [10]. This is the reason behind the constant search for an optimization of the usual US examination. Some studies have therefore approached the tissue characterization, implying the measurement of physical parameters on the US image, in order to obtain a quantification of the ultrasonographic information [11]. However, since the intrahepatocyte fat content leads to an increase in the number of highly reflective interfaces (and therefore to hyperechogenicity) and to an attenuation of ultrasound in the posterior areas, the behaviour of ultrasounds in liver steatosis can also be assessed through the study of attenuation and backscattering coefficients. The backscattering coefficient measures the energy amount returned to the transducer, while the attenuation coefficient (AC) measures the rate of ultrasound absorption according to depth [12]. The aim of this study is to assess how the histological parameters can influence the attenuation coefficient, as well as the performance of this coefficient in differentiating the steatosis degrees in diffuse liver diseases of various biopsy-confirmed etiologies, in comparison to the performance of usual ultrasonography.

II. PATIENTS AND METHODS

A. Patients

The study included prospectively 667 patients admitted to the 3rd Medical Clinic having different diffuse liver diseases: 535 patients having C viral hepatitis (HCV), 36 patients with B viral hepatitis (HBV) and 96 patients having non alcoholic steatohepatitis (NASH). All patients were programmed for a liver biopsy in order to grade and stage their condition. On the day previous to the biopsy, an abdominal ultrasound was performed for all patients, according to a protocol especially designed for the purpose. A series of biological parameters were measured on a blood sample taken after a 12-hour overnight fast: alaninaminotransferase (ALT), aspartateaminotransferase (AST), gamma-glutamyl transpeptidase (GGT), total cholesterol, triglycerides, total bilirubin, basal glycemia and platelets. The study was approved by the Ethics Committee of the University of Medicine and Pharmacy Cluj-Napoca. The nature of the study was explained to each patient and a written informed consent was obtained before enrollment in the study.

Fibrosis detection using ultrasound and serological markers as features for additive logistic regression

Cristian Vicas, Sergiu Nedevschi

Technical University Cluj Napoca, Romania
{cristian.vicas, sergiu.nedevschi}@cs.utcluj.ro

Monica Lupsor, Radu Badea

3rd Medical Clinic Cluj Napoca, Romania
{monica.lupsor, rbadea}@umfcluj.ro

Abstract— Fibrosis and cirrhosis are the main complications of chronic liver diseases. At present, liver biopsy is the golden standard for evaluating the fibrosis. Present study tries to identify novel possibilities for noninvasive fibrosis evaluation. We included 591 patients. From each patient a number of 93 features were obtained using B-mode ultrasound, Doppler ultrasound, transient elastography and serum markers. The patients were grouped in five two-class problems that were presented to two classification schemas, one based on Support Vector Machines and the other based on logistic regression. Cross-validation along with area under curve (AUROC) is used to measure the classification performance. The AUROC of 0.90 was recorded when discriminating between fibrosis stage ≤ 3 and fibrosis stage 4.

Liver fibrosis, non invasive diagnosis, additive logistic regression models

I. INTRODUCTION

The accurate diagnosis and staging of hepatic fibrosis is crucial for prognosis and treatment of liver diseases. To date, liver biopsy remains the gold standard for fibrosis assessment in hepatitis C. Liver biopsy suffers from several important drawbacks. In [1] authors note morbidity, observer variability and sampling variation.

There are several noninvasive alternative methods to investigate the liver: ultrasound investigation including simple B-mode and Doppler mode, current serum markers used to investigate liver function, special serum markers that are involved in fibrosis generation, transient elastography etc.

In [2], using Doppler ultrasonography, authors record the maximum velocity of blood at the portal vein and hepatic artery in 19 patients with cirrhosis, 61 chronic hepatitis and 20 healthy patients. For each patient the arterio-portal (A/P) ratio was computed. The authors noted that the levels of A/P ratio were significantly higher in patients with liver cirrhosis.

Another approach presented in [3] investigates the serum levels of hyaluronic acid, type III procollagen, N-terminal procollagen etc. in patients with hepatic fibrosis. The authors used 114 serum samples from biopsied patients divided in two groups according to their fibrosis stage ($S \geq 1$). They computed an AUROC of 0.8 for procollagen III peptide marker.

The aspartate aminotransferase to platelet ratio index is proposed in [4] as a noninvasive index to predict significant fibrosis and cirrhosis. Using one training set containing 192 patients and a validation set consisting of 78 patients the authors used this index to predict significant fibrosis (Ishak score ≥ 3) and cirrhosis. The AUROC for significant fibrosis in validation set was 0.87 and 0.93 for cirrhosis.

Support Vector Machines along with the sequential forward floating selection scheme was used in [5] to discriminate between patients with no or mild fibrosis ($F \leq 1$) and patients with significant fibrosis ($F \geq 2$). 204

patients were investigated using liver biopsy (METAVIR score) and 34 serum markers. The mild fibrosis lot has 86 patients and the severe fibrosis lot 118 patients. The authors used leave one out cross validation and four selected serum markers (total bilirubin, platelet count, prothrombin time, hyaluronic acid). They achieved an accuracy of 96% in predicting severe fibrosis and 79% in predicting mild fibrosis. The SVM parameters were fixed during the experiments. No special measures were taken to compensate for unbalanced datasets.

Ultrasound based transient elastography is a noninvasive procedure that determines the stiffness of the liver tissue. The liver tissue stiffness is associated with the degree of fibrosis. In [6] the authors perform a review on the performance of transient elastography in detecting liver fibrosis. Their conclusion is that transient elastography is clinically useful in assessing the presence or the absence of cirrhosis. Discriminating between other fibrosis stages is still a challenging task.

In another paper [7] authors review a series of noninvasive tests including various commercially available serum kits, transient elastography and blood flow investigation. Combining biomarkers that measure different components of the fibrosis may advance the diagnostic accuracy. The authors also note that imperfect golden standard might impose some limitations in developing a noninvasive test. This could be compensated by developing longitudinal tests.

In [8] the authors review 153 papers that approach the non invasive fibrosis detection. They included all papers referring to imaging modalities and use liver biopsy as reference. The authors investigate B-mode ultrasonography, Doppler ultrasonography, contrast-enhanced ultrasonography, sonographic elasticity, transient elastography, tissue strain imaging, supersonic shear imaging, magnetic resonance imaging, contrast-enhanced MRI, diffusion-weighted MRI, magnetic resonance elastography, magnetic resonance spectroscopy, computed tomography, positron emission tomography, and single photon emission computed tomography. They conclude that the diagnostic performances of all described noninvasive radiologic modalities were better in distinguishing patients with cirrhosis from lesser degrees of fibrosis. However, staging of fibrosis was rarely achieved reliably. Transient elastography appears to be the most promising technique.

In our previous work [9] we tried to determine a method of detecting fibrosis using image processing algorithms applied on B-mode ultrasound images. In [9] the basic idea was to quantify the liver homogeneity by the means of texture descriptors. The results indicated the possibility of detecting fibrosis by means of non invasive investigations.

In this paper we tried to include most of the noninvasive features analyzed in [1]-[7]. A number of 93 features are acquired from 591 patients. These features include transient elastography evaluation, several common

Ultrasound Assessment Techniques and Quantification of Bowel Inflammation in Several Intestinal Diseases: Initial Experience

M. Socaciu, R. Badea, C. Hagi, S. Iobagiu, A. Golea, L. Ciobanu
 Medical Clinic III
 Cluj-Napoca, Romania
e-mail: socacium@yahoo.com

P. Mitrea, D. Mitrea
 Department of Computer Science
 Technical University of Cluj-Napoca
 Cluj-Napoca, Romania
e-mail: delia.mitrea@cs.utcluj.ro

Abstract— Bowel inflammatory diseases are severe, chronic, reoccurring disorders, whose treatment requires constant efficiency monitoring. Our aim is to find non-invasive methods of assessing inflammation levels. **Material and Method:** we have started a trial by performing clinical, biochemical, B-mode, Doppler and contrast-enhanced ultrasound, together with endoscopy with biopsy on patients with IBD, before and after treatment. Images and time-intensity curves were post-processed using mathematical modeling and texture-based computer analysis. **Results:** early experience (18 examinations) helped us identify characteristic ultrasound features for activity and remission stages of disease: bowel wall layer definition and thickness, wall vessel density and architecture, resistivity index, intensity of layers' contrast enhancement, shape of TICs, TIC parameters and textural GLCM parameters. Further data acquisition is needed to validate those features.

Keywords- Bowel inflammatory diseases; non-invasive diagnosis; contrast-enhanced ultrasound; mathematical modeling; computerized image analysis

I. INTRODUCTION

Bowel inflammation is a common disorder found in several intestinal diseases, ranging from malignant to purely inflammatory ones. Such diseases include: intestinal infections (both bacterial and viral, specific and non-specific), vasculitis and other rheumatologic diseases with bowel involvement, celiac disease and treatment-induced digestive diseases (chemotherapy – neutropenic colitis; radiotherapy – radice enterocolitis; antibiotherapy – pseudomembranous colitis). However, there is a group of intestinal diseases, known as Inflammatory Bowel Diseases (IBD), that have no known causes, but are important because of their severity.

Inflammatory bowel diseases are chronic diseases of the digestive tube (mainly Crohn's disease –CD– and Ulcerative colitis –UC–), characterized by an inflammatory process of the intestinal wall. The inflammation extends to all the layers, including the mezentery and the intra-abdominal lymph nodes, in Crohn's disease and only to the mucosa in the Ulcerative colitis. Patients with inflammatory bowel diseases present alternatively periods of remission and of activity during which there are numerous symptoms and signs. During their lifetime, patients need several hospitalizations. The evolution is marked by the onset of complications, among which some of major severity, so that the assessment of the inflammation activity and of the treatment response is crucial in monitoring these patients.

There are several clinical, laboratory and paraclinical parameters used to assess the activity phase. Most used clinical scores are the Crohn Disease Activity Index (CDAI) and Truelove Witts (for UC). Together with laboratory parameters like: erythrocyte sedimentation rate, C reactive protein, clearance of stool α_1 antitrypsin, they can assess to some degree the activity but are not enough accurate.[1]

The standard methods of diagnosis and of inflammatory phase assessment are the endoscopic, radiologic and histopathology exams. But these are too invasive for severe forms and they cannot be permanently repeated in order to monitor the clinical evolution. Also, they only provide information from the mucosal layer, with the other structures being inaccessible.

Computer tomography and magnetic resonance imaging are elective imagistic methods used to identify the intra-abdominal and intestinal wall complications, but are less accessible and cost a lot. Ultrasonography has similar potential in diagnosis, but with advantages like: noninvasivity, reduced cost and the possibility of repeatability. Numerous literature studies have proven its role for the examination of digestive tube pathology.

Our aim is to develop new methods of activity assessment in IBDs, based on ultrasonography examination, combined with some modern techniques like vascular contrast enhancement and computer-aided analysis of images.

II. MATERIAL AND METHODS

Our study takes place in Medical III Clinic, Cluj-Napoca, since march 1st 2009. All the clinical and paraclinical exams are performed in our clinic, while the database and computerized analysis are developed at the Department of Computer Sciences from the Technical University, Cluj-Napoca.

Patients already diagnosed with Intestinal Bowel Disease, with positive endoscopy exams, were recruited into the study. The patients were excluded from the study in case they were in poor condition (unable to hold breath, for example), with a surgical procedure on the intestine or if they didn't consent with the examinations.

For each patient, a full set of investigations is done at first admittance (before treatment) and after a period of treatment (typically 1 month). Investigations include: clinical exam, biochemical exam, endoscopy with

Improving the imagistic textural model of the hepatocellular carcinoma through dimensionality reduction techniques and classifier combinations

Delia Mitrea, Sergiu Nedevschi
 Computer-Science Department
 Technical University of Cluj-Napoca
 Cluj-Napoca, Romania
Delia.Mitrea@cs.utcluj.ro

Monica Lupsor, Mihai Socaciu, Radu Badea
 Department of Ultrasonography
 3rd Medical Clinic
 Cluj-Napoca, Romania
mmlupsor@yahoo.com

Abstract— The non-invasive diagnosis of the malignant tumors is a major issue in nowadays research. We aim to elaborate computerized, texture-based methods, in order to perform automatic recognition of the hepatocellular carcinoma, which is the most frequent malignant liver tumor, using information from ultrasound images. We elaborated previously the imagistic textural model of HCC, consisting in the exhaustive set of textural features, relevant for HCC characterization, and in their specific values for HCC. In this work, we improve the imagistic textural model and the classification process, through dimensionality reduction techniques, focusing on the feature extraction methods, and classifier combinations. From the feature extraction methods, we implemented the most representative one – PCA, together with its extension, Kernel PCA [16]. Concerning the classifier combinations, the voting, bagging and boosting procedures were compared. These methods were evaluated in order to accurately distinguish HCC from the cirrhotic parenchyma and from the benign tumors.

Keywords - liver tumors, Principal Component Analysis, classifier combinations, improvement of the classification process

I. INTRODUCTION

The hepatocellular carcinoma (HCC) is the most frequent malignant liver tumor (75% of liver cancer cases), besides hepatoblastoma (7%), cholangiocarcinoma and cystadenocarcinoma (6%). The human observations are not enough in order to give a reliable diagnosis, and the biopsy is an invasive, dangerous method. Thus, a more subtle analysis is due, and we perform this by using computerized methods, in order to extract the necessary information from ultrasound images. The texture is a very important visual feature in this context, as it provides a lot of information concerning the pathological state of the tissue; it describes the regular arrangement of the grey levels in the region of interest, being also able to provide multi-resolution parameters. The texture-based methods, in combination with classifiers, have been widely used for the automatic diagnosis of various kinds of tumors [1], [2], [3], [4]. However, a systematic study of the relevant features and of their specific values for the characterization of HCC, based only on information extracted from ultrasound images, of the possibilities of obtaining an optimal imagistic model of HCC, and best results for classification, is not done yet. We aim to do this in our research, which consists in modeling the HCC tumor and the visually similar tissues through textural features, and also in analyzing the possibilities of classification process improvement, through feature extraction and feature selection techniques, as well as through classifier combinations. In this work, we analyze the effect that the dimensionality reduction techniques and classifier combination procedures can have on the improvement of

the classification process. We focus on the feature extraction methods, which improve the speed and accuracy of the classification, by reducing the dimensionality of the data, while emphasizing the important characteristics. The method of Principal Component Analysis (PCA), is the best known linear feature extractor [5], which aims to represent the data into a new space, lower in dimensions, by finding the best projection directions, which minimize the distance from the data points in the least-square sense, and also emphasize the data variability. The Kernel PCA method is a generalization of PCA, which maps the data from the initial space, into a space that is appropriate for PCA representation [6]. Another approach, which gave good results concerning the improvement of the classification process, is the implementation of the classifier combination procedures. We compared the voting technique, as well as the more complex bagging and boosting combination modalities, which involved also data resampling and hierarchical structuring of the classifiers. As weak classifiers, we used the methods of Multilayer Perceptron, Bayesian Belief Networks and Decision Trees, that gave best results in our previous experiments [9].

II. THE STATE OF THE ART CONCERNING TUMOR CLASSIFICATION BASED ON MEDICAL IMAGES

The most frequently used methods in the field of texture-based characterization of the malignant tumors are the Grey Levels Co-occurrence Matrix (GLCM) and the associated Haralick parameters, the Run-Length Matrix parameters [1], fractal-based methods [2], the Wavelet [3] and Gabor transforms [4], combined with the k-nn classifiers, Bayesian classifiers [2], Artificial Neural Networks, Fisher Linear Discriminants [1], Support Vector Machines [3]. In [1] the authors compute the first order statistics, the Grey Level Co-occurrence Matrix and the Run-Length Matrix parameters, which are used in combination with Artificial Neural Networks, as well as with Linear Discriminants, for the classification of the liver lesions. The fractal-based methods and the multiresolution methods are more emphasized, because of the chaotic, complex structure of the tumor tissue. The fractal-based methods are used in [2] in order to distinguish the salivary gland tumors from ultrasound images. The Wavelet transform was also implemented in order to analyze the values of the textural parameters at multiple resolutions, for differentiating malignant and benign liver tumors from ultrasound images [3]. Concerning the implementation of the dimensionality reduction methods, Decision Trees [10], Genetic Algorithms and Neural Networks [11], as well as Independent Component Analysis (ICA) combined with Least Square Support Vector Machines (LS-SVM) [12], were applied successfully. The classifier combinations were experimented in order to distinguish the human brain tumors within magnetic resonance images [13]. A voting

Abdominal Lesion Segmentation within CT images. A survey

Anca Ciurte and Sergiu Nedevschi
 Technical University of Cluj–Napoca

Email: Anca.Ciurte@cs.utcluj.ro, Sergiu.Nedevschi@cs.utcluj.ro

Abstract

Lesions segmentation is a difficult task in the analysis of medical images. Fully automated methods happen to fail, producing incorrect results and requiring the intervention of a human operator. Segmentation in medical imaging domain is particularly difficult, due to the noise and low resolution of the medical images caused mostly by the acquisition techniques and also by the high biological and pathology variation. This paper reviews the latest approaches in abdominal lesion segmentation within CT images. A number of twenty one papers were studied and compared.

Index Terms: image segmentation, abdominal lesions, computed tomography, review.

1. Introduction

Lesion segmentation is a key task in many clinical applications such as localization of pathology, treatment planning, radiotherapy, lesion surveillance in time, computer-integrated surgery, cryo-ablation and others. In the process of medical intervention planning, the decisions are always taken based on medical imaging technology including computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET) or ultrasound (US). The main drawbacks in processing medical images are related to their quality, medical images being characterized by low contrast and heavy noise. These drawbacks are mostly caused by the acquisition techniques, by the patients movements and by the organs movements (due to breathing and heart beating). Among these imaging techniques, CT is often preferred by diagnosticians since it provides an accurate information about the anatomical structure thanks to its higher signal-to-noise ratio and better spatial resolution, providing detailed 2D sections with information on 12 bits.

Segmentation of CT images is a difficult task, especially when talking about lesion segmentation which brings more challenges because of the large variety of lesion types, density, contour, shape and others. Most often, the radiologists use semi-automatic segmentation techniques, their intervention being required to select the areas of interest and/or to manually initialize the segmentation algorithm. The robustness and reliability of these semi-automatic segmentation techniques is demonstrated, compared with manual segmentation which is established to be the *gold standard*. Despite that, automatic segmentation techniques are encouraged due to the major advantages that can be brought such as the volume of information handled, the required processing time, the required experience of the user and the repetitiveness nature of the task.

The purpose of this survey is to categorize and briefly review the literature on computer aided lesions segmentation within abdominal CT scans, with an emphasis on the techniques that have been employed. The paper has the following structure: the first section presents an introduction into the medical imaging domain concerning the problem of lesion segmentation in CT images; the second section briefly presents the abdominal lesion segmentation approaches in the medical imaging domain, categorize them as following: thresholding, region growing, watershed, graph cuts, active contours, level sets, statistical algorithms, neuronal networks, machine learning and random walker based approaches; the second section contains also a review over the evaluation techniques for the lesion segmentation algorithms; the third and the final section presents some discussions and conclusions.

2. Abdominal Lesion Segmentation Techniques and Validation Measures

In the following, the latest segmentation techniques used for abdominal lesions delineation are described. Twenty one papers were studied, all published in the medical imaging domain in the last decade. *Table 1* summarizes the studied techniques by presenting basic information like the type of method, the dataset, the type of lesions and the studied organ, and also some results that evaluate the method performance. A common classification of these lesion segmentation techniques is as follows: thresholding, region growing, watershed, graph cuts, active contours, level sets, statistical approach, neural networks-based methods, machine learning and random walker.

A global structure of the lesion segmentation techniques is presented in Fig. 2.1. Generally some preprocessing steps of image enhancement are performed to reduce the noise and increase the image contrasts, selection of a region of interest in order to reduce the task complexity and also the initialization of the segmentation algorithm. For ROI selection several approaches exist. Some perform a segmentation of the studied organ, a task that also encounters many difficulties. Others, manually mark a ROI (more often rectangular shapes) and impose constraints that will affect the lesion segmentation process (e.g. the rectangle center is the same with the center of the tumor, the ROI volume is proportional with the tumor volume, etc.). Depending on the used segmentation algorithm and their specifications, these preprocessing steps might be missing. Next, the image segmentation is performed with the aim of realizing the lesion delineation using variate segmentation techniques, as presented in the following. This step is often followed by a postprocessing step for the refinement of the lesion contour.

Complex Histological, Ultrasonography and Clinical Studies in early noninvasive diagnosis of the Photo-induced Cutaneous Senescence and in the Photo-induced Skin Cancers, using Computerized Imaging, Modern Biotechnology and Mathematical Modelling Methods

M. Crisan¹, C. Cattani², R. Badea¹, R. Cosgarea¹,
S.Dudea¹, M. Lupsor¹, A. Stir¹

¹University of Medicine and Pharmacy "I Hatieganu"
Cluj-Napoca, Romania,

²University of Salerno, Salerno, Italy
e-mail: mcrisan@yahoo.com

P. Mitrea, D. Mitrea

Department of Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
e-mail: paulina.mitrea@cs.utcluj.ro

Abstract— This article subscribes to the most modern research line - based especially on computerized imaging, of both the photo-induced cutaneous senescence and the photo-induced skin cancer. The ultra-violet rays represent the major factor involved in the cutaneous senescence, by determining molecular, cellular, tisular and clinical changes. Our target is the scientific fundamenting and the development/standardization of new noninvasive skin investigation and diagnosis methods, like High-Resolution Ultrasonography and Age-Reader (Glycated-collagen). The noninvasive assessment of the glycated collagen degree in the skin, quantifies the skin age/senescence degree and rate of instalation/risk for the development of age related skin pathologic conditions or other chronic diseases and their complications, including skin cancer. The elaboration of mathematical models of basal cell and spindle cell carcinoma development, and their software implementation, is based on correlations between specific molecular markers of each tumoral type, their morphological and ultrasonography aspects and clinical evolution.

Keywords - skin cancer, computerized imaging, High-Resolution Ultrasonography, Age-Reader, non-invasive diagnosis

I. INTRODUCTION

The significant increase in life expectancy and the process of population aging are aspects that generate important social and economic changes and influence the health and research policies throughout the world. The senescence phenomenon represents a natural, slow and irreversible process, which affects all body tissues, being determined by a multitude of factors that contribute in different proportions to the characteristic molecular, cellular, tisular and clinical changes[1].

The definition of standard objective parameters, appropriate to different biological age groups, is needed in order to determine the contribution of different factors to the cutaneous photoaging; the comparison between individual measurements and these standard parameters allows: the assessment of the degree and installation rate of the senescence process, the identification of persons at risk for developing premature aging associated pathology, suitable therapy for the slowing down of the processes implicated in senescence.

The tegument is a biomembrane situated at the interface with the external environment, continuously exposed to the aggression of the actinic factor; it reflects the state of health of the body and the human personality and has numerous psycho-social implications. Several scientific theories on the aging regulation at molecular, cellular and systemic levels have been postulated in order

to define and control this process. Aging has genomic internal causes that are activated by exogenous or epigenetic factors. The **ultra violet rays** represent the major factor involved in the process of cutaneous senescence[2].

II. THE MOST KNOWN THEORIES IN THE DOMAIN

The theory of error accumulation described by Orgel in 1963 interferes with the DNA replication, RNA transcription and translation into proteins. The risk for errors increases with age; this determines at a certain moment a critical change in the genome that triggers senescence, apoptosis and cellular death[3].

The free radical theory postulated by Harman in 1956, demonstrated that the reactive oxygen species interfere with the cellular and subcellular systems, thus inducing molecular degradations.

The mitochondrial theory is based on the fact that the mitochondria is the main source of free radicals within the cell. Skin exposure to the sun determines the accumulation of mutations in the mitochondrial DNA, with implications in senescence.

The theory of protein glycation or the **cross-linking theory** focuses on another potentially destructive agent, **glucose**. This binds to the proteins in the extracellular matrix, including the collagen, during non-enzymatic reactions generally called glycation and forms end products of glycation called AGEs (advanced glycation end products). The UV rays determine the accumulation of AGEs in the elastic and collagen fibers in the basal membranes, dermis, vascular walls, with the alteration of the physical and mechanical properties of the extracellular matrix[4]. The fluorescence of the **tryptophan** in the **epidermis** and **glycated collagen in the dermis** are considered in vivo **objective markers** of the aging process. The glycation end products have been first described in diabetes patients. Because of the numerous analogies between the effects of diabetes mellitus and those of the senescence process, diabetes is considered a model of accelerated aging[5]. Literature data show that diabetes is associated with cutaneous pathology with certain prevalence on different age segments: juvenile acne, pityriasis versicolor, verrucae, epidermal cysts, microbial and mycotic infections especially in young patients, cutaneous xerosis, solar elastosis, seborrheic and actinic keratoses, seborrheic dermatitis, cutaneous tumors, pigmented macules in the case of adults and elderly people[6].

The A and B ultra violet rays induce and accelerate the glycation process, interact with cellular and vascular components, induce the synthesis and release of cytokines,

Discretization Methods for obtaining Energy-Minimizing Deformable Models

Alexandru I. Mitrea ,Octavian Mircia Gurzău

Department of Mathematics
Technical University of Cluj-Napoca
Cluj-Napoca, Romania

Alexandru.ioan.mitrea@math.utcluj.ro, gurzau@math.utcluj.ro

Abstract— This paper is devoted to characterize some energy-minimizing deformable models by means of variational principles and to obtain numerical algorithms in order to detect the contours, which perform the minimum of the associated energy-functional. The graphs of some 2D models of this type are given, too.

Keywords: energy-functional, isotropic deformable model, finite difference discretization, evolution equations, pentadiagonal matrix

I. INTRODUCTION

The theory of the deformable models has appeared, in computer vision, in the mid of 1980's and it originates from the researches of D. Terzopoulos regarding the theory of multidimensional deformable models, [9], [10]. The papers of D. Terzopoulos, A. Witkin, M. Kaas [12] and D. Terzopoulos, K. Fleischer [11] are recognized as the beginning of the researches concerning deformable models in computer vision and computer graphics, respectively; the deformable model that has attracted the most attention (the active contour model) is popularly known as "snakes", [4]. Snakes are plane deformable curves that are useful to solve the problem of object segmentation, an important task in medical image analysis; in fact, the deformable models represent a vigorously researched model-based approach to computer-assisted medical image analysis, [5]. The shape of a contour subject to an image $I(x,y)$ is determined by the so-called energy-functional $E(v)$, which is, generally, a functional of integral type; in order to find the contour which minimizes $E(v)$ principles of calculus of variations and numerical algorithms are used. In fact, the mathematical foundations of the deformable models involve various mathematical fields (Functional Analysis, Approximation Theory, Differential Equations, Differential Geometry, Numerical Methods, Calculus of Variations).

Starting from the original model of D. Terzopoulos, M. Kaas and A. Witkin [12], various deformable variational models were proposed, such as: the Balloon-Models (I. Cohen and L.D.Cohen, [3]), the Adaptive-oriented g-Snakes Model (F. Preteux and N. Rougon, [8]), the Interpolation Snakes (S. Minut and G. Stockman, [6]), the Multiscale Active Models (X. Bresson, P. Vanderghiest and J.P. Thiran, [2]). Let us remark, also, the contour detection performed by means of dynamic programming of A.A. Amini et al. [1], with applications to echocardiographic images, [7].

The goals of this paper consist of finding the differential vectorial equation which provides energy-minimizing contours with respect to an isotropic 2D model, obtaining the matricial system corresponding to the discretization of the previous equation, solving the differential equation or the matricial system and, consequently, finding the energy-minimizing contour.

The next section offers a description of the 2D isotropic model from mathematical point of view. The third section is devoted to derive the (ELP) Equation associated to this model. Two examples, together with the corresponding graphs, are given, too. The fourth section presents the numerical algorithm obtained from (ELP) Equation by discretization with finite central differences; a numerical example and the corresponding graph illustrate the theoretical formulas. Finally, the dynamic isotropic model, based on the method of evolution equations is inferred, together with the corresponding numerical algorithm.

II. MATHEMATICAL DESCRIPTION OF THE DEFORMABLE MODEL

The position of a deformable active 2D model (snake) is given, geometrically, by a parametric curve:

$$v = (x, y)^T; \quad v(s) = (x(s), y(s))^T, \quad 0 \leq s \leq 1 \quad (\gamma)$$

where x and y are real function of class $C^2[0, 1]$, i.e. v is a vectorial function of class $C^2([0, 1], \mathbb{R}^2)$.

Let $\mathcal{A} \subseteq C^2([0, 1], \mathbb{R})$ be the class of admissible deformations, which consists of all deformable curves $v \in C^2([0, 1], \mathbb{R}^2)$ so that the values $v(0)$, $v(1)$, $v'(0)$, and $v'(1)$ are given. In order to find the optimal position of the snake, it is necessary to characterize its state, by means of the corresponding **energy-functional**. The energy of a snake is given, usually, by the sum of the internal energy, external energy and, optionally, the energy associated to some constraint forces. Further, let us define the energy functional of a snake. To this purpose, the following data will be introduced.

- two scalar nonnegative functions, $\alpha(s)$ and $\beta(s)$ of class $C^2[0, 1]$, named also **weight-functions**, which control the elastic and the bending properties of (γ) , respectively
- a real function $I(v) = I(x, y)$ of class $C^2(\mathbb{R}^2)$, named **image intensity**,
- a real function $P(v) = P(x, y)$ of class $C^2(\mathbb{R}^2)$, which represents the **potential associated to the external forces**,
- a vectorial function $k(v) = k(x, y)$, $k(x, y) \in C^1(\mathbb{R}^2, \mathbb{R}^2)$ which refers to the local dilatation or local contraction of (γ) along its normal; usually, $k(v) = cv$, with $c \in \mathbb{R}$.

The **energy-functional** associated to these data is defined by the equality:

$$E(v) = E_{int}(v) + E_{ext}(v) + E_{balloon}(v) \quad (1)$$

The **internal energy**

$$E_{int}(v) = E_{els}(v) + E_{bend}(v) \quad (2)$$

is obtained by adding the **elastic energy**

Computer-aided Remote Patient Monitoring with Personal Healthcare Units

Gheorghe Sebestyen, Lorand Krucz
 Technical University of Cluj-Napoca
 gheorghe.sebestyen@cs.utcluj.ro

Gavril Saplacan, Mircea Rusu
 Company for Applied Informatics
 office@cianet.ro

Abstract— In order to allow real-time monitoring of patients and also to provide means of intervention in remote locations over the Internet a set of modular Personal Healthcare Units (PHU) where developed. These units are wearable smart medical devices that measure critical medical signs of a patient and transmit them to a healthcare center.

A specialized communication protocol was developed in order to connect these devices through a wireless connection to the CardioNet distributed medical system. This approach improves healthcare quality through continuous supervision and also assures mobility of patients with cardio-vascular diseases.

I. INTRODUCTION

One of the main objective of the “i2010” EU policy framework (towards an „Information society”) is the extensive use of IT&C technologies in healthcare systems in order to improve the quality of medical services, increase the system’s responsiveness and reduce costs[1]. This IT&C approach will impose fundamental changes in the way healthcare is delivered and medical knowledge is managed and transformed into clinical practices. Recently a significant number of national and international eHealth projects were initiated:

- Projects which strive to find a common platform for medical data interchange, through medical communication protocols (e.g. HL7 [5])
- Projects concerned with classification and unique coding of medical concepts, terms, diseases, drugs and even medical procedures [3,4,6]
- Projects concerned with medical data-mining, reasoning, statistical processing - used as support for improved medical diagnosis and treatment.
- Projects developing wearable low-cost healthcare units to improve the quality of life for chronic patients and encourage home healthcare.

A modern approach to medical service provisioning must be a consumer centered approach. Patient must be able to receive quality medical services at their convenience. It constitutes an important factor of convenience if the access to medical services does not require the patient to break his or her daily routine and spend large amounts of time in medical facilities under observation, but rather medical facilities integrate into his/her daily life. This integration is made possible by remote monitoring and control of patient health via small portable medical devices called Personal Healthcare Units. Integrating these units into a large scale distributed medical application through networking allows real-time and continuous access to comprehensive medical services at any time and from any location.

The development of Personal Healthcare Units was initiated as part of the CardioNET project, which is meant to offer a patient-centric and distributed medical service framework, dedicated for patients with cardio-vascular diseases. The main objectives of the CardioNET project are:

- Evaluation of the latest achievements and solutions in the eHealth field
- Analyses of the legal background for IT&C technology use in health systems
- Definition of a generic framework for tele-cardiology applications
- Implementation of a distributed software system that assures remote monitoring of patients and seamless access to medical services
- Definition of a flexible Electronic Health Record (EHR) system
- Development and implementation of medical applications specialized for different medical entities: general practitioner, hospitals, medical laboratories, etc.
- Definition of a web-service based interface between different medical applications that facilitate seamless exchange of medical data
- And last but not least, development of portable medical devices for continuous monitoring and data acquisition

II. THE CARDIONET DISTRIBUTED MEDICAL FRAMEWORK

The CardioNET project [13,14,15] is an interdisciplinary applicative research project focused on improving medical services and also reduce overall healthcare costs through IT&C technologies. The CardioNET framework is designed to comply with national and international standards and trends in cardiology and it is implemented in a flexible and modular way that allows continuous extension and updates. Even if it was initially meant for cardiac patients, with small extensions it may be used for supervision of other chronic diseases. In order to assure interoperability and data exchange with different medical applications the CardioNET system is built upon a domain ontology called the *heartfaid ontology* [2].

The framework consists of 3 types of nodes:

- The generic CardioNET medical facility application which is a web application intended for use in medical facilities such as medical cabinets, clinics hospitals etc..
- The home healthcare personal area network, which is intended for home use by patients
- The CardioNET repository and portal, which is a central node meant to provide general medical assistance and access to medical services

The generic CardioNET medical facility application contains 3 major elements:

- A database, which stores the general medical knowledge base, the ontology and the specific medical data of patients that are or were treated in the specific medical facility.

Software solution to quality assessment in medical analytical laboratories

Lucia Vicas, Diana Pelea

University of Oradea
Romania,
{lucia_vicas,diana_pelea}@yahoo.com

Cristian Vicas

Technical University Cluj Napoca
Romania
cristian.vicas@cs.utcluj.ro

Abstract—Quality control and quality assessment in medical analytical laboratory ensures that the various analytical measurements performed have a high degree of accuracy, precision and reproducibility. The most used tools in quality assessment are the Levey-Jennings charts along with Westgard rules. In this paper is presented a software solution that implements the most common Westgard rules and ensures the autoimmunization of routine quality assessment. The basic quality control notions along with corresponding statistical fundamentals for the Westgard rules are also presented.

Quality control, software solution, westgard rules

I. INTRODUCTION

Quality control is a procedure or set of procedures intended to ensure that a manufactured product or performed service adheres to a defined set of quality criteria or meets the requirements of the client or customer.

Quality assurance is defined as a procedure or set of procedures intended to ensure that a product or service under development (before work is complete, as opposed to afterwards) meets specified requirements. Quality assurance is sometimes expressed together with quality control as a single expression, quality assurance and control (QC) [1].

In medical laboratory these terms are applied to various measurements performed on human fluids. There are a varied range of substances measured, using various methods. Most of the procedures determine the concentration of a solution by measuring the quantity of absorbed light at certain wavelength. Others, measure other properties. However, at some point, every method implies an electric measurement. This measurement is affected by Gaussian noise. In addition to this noise, there are other noises, like thermal noise, impurities in the substances, wrong dilutions of the reactants, etc. Some of these sources of noise can alter the results in a clinically significant amount. Quality control tries to maintain a low level of noise into the analysis process. Ideally it will detect any noise that has other source than implicit Gaussian noise.

In order to perform a quality control, physicians follow a specific procedure: A special designed serum named control serum with known concentration is analyzed using regular procedures (i.e. as if it comes from a patient). Its values are recorded on a special chart, the Levey-Jennings chart. The measured value for the control serum is called the control value.

Using basic statistical tools these values are analyzed. When the recorded values are not in specific bounds, or they violate certain rules, the physician declare the current state of QC as “out-of-control”, meaning that the analysis returned by the device are not reliable and actions must be taken in order to correct this problem. After the problem is

corrected all the analysis performed on the patients since the last validated control are performed again.

Sometimes it is recommended that several control serums are used for a certain substances (i.e. one control serum is set at a clinically normal concentration and another serum at a clinically abnormal concentration). If for a substance we use two or more control substances of different concentration, we collect a control value for each concentration.

The operation of determining the control values is named a run, or performing a control run.

Some analysis require a very high degree of accuracy and precision because the measured quantities are in range of 10-6 g/ml. (i.e. hormone levels or various cancerous markers) A tight quality control is required to maintain accurate and precise results [2].

For each substance that the laboratory can measure a different quality control protocol must be established.

Manual handling of quality control data is cumbersome and error prone. There are several commercial software available, but they lack the flexibility of defining complex quality control rules. In this paper is presented a software system that enables a high degree of flexibility in defining the control rules and in defining the quality control protocol.

We also give comprehensive statistical notions on which the Westgard rules are based.

The paper is organized as following: in Chapter 2 are presented the most used Westgard rules along with their statistical foundations, in Chapter 3 is presented the proposed software solution and in Chapter 4 conclusions to this paper are drawn

II. WESTGARD RULES

Each measurement is subjective to a Gaussian noise. For a series of measurements of the same physical quantity we obtain values that are distributed according to a Gaussian law. A Gaussian distribution is governed by two parameters, the mean and the standard deviation. In Equation 1 is shown the Gaussian distribution law. For an observation, this function gives the probability of measuring a certain value.

$$p(x) = \varphi_{\mu, \sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (1)$$

where μ is the mean of the distribution and σ is the standard deviation.

Let x_i be a series of N observations that follows a normal distribution. Then,

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i, \quad (2)$$

“e-Romania” from Concept to Implementation

Doina Banciu

National Institute for Research and Development in Informatics,
ICI, Bucharest
doina.banciu@ici.ro

Abstract— The paper presents the concepts and implementation for the portal e-Romania. The “e-Romania” concept, launched in June 16, 2009 by the Romanian Ministry of Communications and Information Society is the mandatory and necessary starting point for the development of e-Romania Portal. For testing purposes the configuration of e-Romania portal was based on a Content Management System (CMS).

I. INTRODUCTION

„An information society for all” has represented a European motto since as early as 2002, when the action plans “eEurope”, “eEurope+” (the latter was dedicated to EU accession candidate countries at that time) and “e-Europe 2005” were launched. In the same period, the Lisbon strategy, re-launched in 2005 within the European Council, was aimed to make the EU a strong and competitive knowledge-based economy by 2010.

The 2010 strategy had three main priorities:

- Creation of a single European information space meant to promote an open and competitive digital economy that may capitalize on the European market and services related to the information society;
- Support for innovation and investment in IT&C research in order to promote the economic development;
- Creation of a comprehensive information society which will ensure the economic and work force development that will further lead to better quality public services and better quality of life;

Of course, nowadays reality does not fully meet public expectations; however, the digital information has undoubtedly gained an unprecedented leading role in the development of all domains of activity as well as of each individual.

In the cultural sector, the access to digital information has always been unanimously recognized as a major influencing factor in the personality development as well as in the modern workforce development and competitiveness.

The initiative to create a single point of access to digital cultural information at the European level has also been taken by the European Union that promoted and supported the European Digital Library also known as “Europeana” by means of various projects.

The European Digital Library [3] has been run by the EDL Foundation and includes outstanding holders of cultural information. Europeana has been developed by cooperation within a “thematic network” named *EDL net. (The European Digital Library Network)*.

The European partners that take part in *Europeana* are institutions from various categories: national archives administrations, audio-video collection holders (such as

Association des Cinemateques Europeenes, European Broadcasting Union, TV2, Danmark etc.), partners from interdisciplinary fields (Museums, Libraries and Archives Council UK, Institute National de l’Audiovisual, etc.), libraries (both national and other categories of libraries that hold collections), museums, research consortia established as a result of the development of some projects such as DELOS, DISMARC, DIGMAP, research & development organizations, appointed national representatives for the library sector. Each participant is liable for the digital information provided for *Europeana*.

As early as the end of the past millennium, when the EU concept “e-Government” or new technology-based government was launched, central and local public administrations from various countries started to create suitable information tools for the citizens to find what they were looking for. These tools have been developed on products and services specific to digital information.

Starting with the widely known “info-kiosks” which offer information services and even ticket-issuing in various public places and ending with interactive portals, the relationship between citizen and administration is frequently based on digital services. As was the case with the digital cultural information, the wide spreading of information has triggered some undesirable side effects, out of which the most important is probably the citizen’s feeling of “getting lost” in the tremendously large volume of digital information often accompanied by “the information noise”.

Hence, various initiatives were taken in many countries of the world with a view to creating single access points to digital information at national level (e-Austria [4] initiative or e-Mexico Portal [5], etc.).

II. E-ROMANIA CONCEPT

In Romania, in the recent years, the public administration, libraries, museums, tourism offices, church organizations, etc. have developed their own web sites which comprise a significant volume of digital information. The previously mentioned effects of “getting lost and informational noise” have started to show up frequently for the Romanian users as well when they look for information in the Romanian language. Therefore, setting up a unitary portal of single access to the digital information in Romania represents an initiative that has been necessary in Romania for so long.

The “e-Romania” concept, launched in June 16, 2009 by the Romanian Ministry of Communications and Information Society is the mandatory and necessary starting point for the development of *e-Romania* Portal. The creation of *e-Romania* Portal proves to be an elaborate work which entails significant time and resources as well as relevant professional competence on at least two levels:

Integrated On-line System for Management of the National Retrospective Bibliography - SIMBNR

Dora Coardoş , Vasile Coardoş, Cornelia Lepădatu

National Institute for Research and Development in Informatics – ICI, Bucharest
coardos@ici.ro, vasile@ici.ro, cornelia@ici.ro

Abstract— The paper presents the results of research obtained in the management of national retrospective bibliography. It describes the design of the on-line integrated system for management of national retrospective bibliography. It also presents the content of Romanian works' bibliography, as well as, the structuring criteria. Finally, it presents the bibliographical information portal attached to the SIMBNR on-line system

I. INTRODUCTION

In the i2010 initiative, it provides for the creation, promotion and support of the European Digital Library (EDL), by associating digital libraries, extant or being created in countries members of the European Union. Thus, there are taken into consideration stimulation of digital content creation, improvement of preservation and enlargement of the public access to collections organized in digital manner.

Transposed into electronic format, the cultural, scientific and educational heritage can become a resource for a large specter of informational products and services, in sectors such as education, research and tourism.

The national bibliography of any country represents the cultural, spiritual and economic evolution of that specific country, a mirror of achievements of all domains, recorded in writing.

For valorization and dissemination of particularly valuable information, contained within the old and modern Romanian literature, it is necessary to create an on-line consultation alternative by using multimedia technologies, in order to obtain a more rapid access of the user, based on structuring criteria of bibliography content. The Romanian literature national retrospective bibliography – work of national interest – describes and systemizes texts that were printed in the Romanian territories representing all books published in Romanian language regardless of the author and location, and also, all works of Romanian authors regardless of language and publishing location during the 1508 - 1918 periods. It is structured, historically, as follows:

- **Old Romanian Bibliography (ORB)**, representing the 1508-1830 period, contains books published in Romania or outside but in Romanian language, or by Romanian authors in other languages. It uses the ISBD(A) as description standard; works are described chronologically, and can be found at the end of every volume in an alphabetical index [6].
- **Modern Romanian Bibliography (MRB)** [7], representing the 1830-1918 period, chronological continuation of Old Romanian Bibliography, contains all books printed in Romanian territories, all books in

Romanian language, regardless of author and publishing location, and all books by Romanian authors, regardless of language and publishing location. The Bibliography is realized according to ISBD(M) standards, mainly based on book fund from the Romanian Academy Library, and is structured alphabetically by author and titles.

The work elaborated and published over time by the Academy Library, in small press run, and presently, out of print, there still are in the library a small number of copies of the printed volumes and records in digital format on magnetic media.

In this context, we intended to design a software instrument that would create a clear image, actual and accessible on-line of the Romanian literature national retrospective bibliography. The content, structure and information retrieval criteria will be defined according to standards library, complying with the European and in the same time national specificities. The system will be an instrument at hand for the experts in biblioteconomy, but also, it will come in handy for other researchers preoccupied by the history of Romanian science, culture and art.

Improvement, coordination and optimization of the process of retrieval of information contained in the Romanian literature national retrospective bibliography are based, mainly on two activities:

- identification of structuring criteria for the content of the old and modern Romanian literature bibliography (national retrospective bibliography)
- creation of mechanisms for coordinating, managing and maintaining the process of information retrieval that will constitute the digital content of the national retrospective bibliography, in order to allow a coherent architecture for the entire system.

Based on information gained during the process of digitizing the two types of bibliography - Old Romanian Bibliography (ORB) and Modern Romanian Bibliography (MRB) - components of the Romanian Literature National Retrospective Bibliography, the integrated system is dedicated to the management, retrieval and valorization of this cultural information fund. In achieving these functions, two main components can be observed in the system's architecture:

- the component for creating, loading and managing the bibliographical database
- the component for on-line retrieval, based on several criteria, and consultation of bibliographical information using multimedia technologies

Ontology-based content and learning personalization in an e-learning system

Ana-Maria Borozan, Ileana Trandafir

National Institute for Research and Development in Informatics
(marika, itran)@ici.ro

Abstract— The paper presents the framework and the advantages of the learning content development based on ontologies (domain concepts and competences), learning objects and metadata. Applying these new concepts a personalized learning path - a chain of concepts and his associated learning objects - may be composed for each learner in conjunction with his objectives, cognitive state and preferences. The original personalization model, the learning content and the e-learning tools developed within CEEEX project “Innovative System for Personalized and User-centred Learning with Application to Project Management” - SinPers, demonstrates the feasibility and the opportunity of this approach, integrating the content management with unit of learning personalization and execution functions.

I. PERSONALIZATION - AN ADVANCED APPROACH IN THE E-LEARNING SYSTEMS

In the near future lifelong learning will become a new form of work; the use of knowledge acquired in school is made at the working place, and the professional activity is more and more relying on intensive-knowledge. The learning will be ubiquitous and permanent both for individuals and organizations, in the formal training, in professional environment or at home, including entertainment applications. The e-learning systems must become accessible to each citizen, regardless age, education and social statute, therefore, as adaptable to its needs as possible, independent of time, place and pace. As consequence, services and content personalization offered to the users is an advanced phase in the e-learning systems evolution and also a priority of the research in this domain [1].

Thus the 7th European Framework Program, section “Technology-enhanced learning” proposes the following priorities: “the creation of responsive environments for technology-enhanced learning and, at longer-term, the development of adaptive and intuitive learning systems. In particular, they should be able to accommodate *personalisation* to respond to specific learning needs and contexts and enhance *competence*, skills and performance”.

The web offers a perfect technology and environment for individualized learning because the learners can be identified uniquely, the provided content can be personalized and the learners’ progress can be traced, assisted and evaluated. The Semantic Web concepts carry a new way of semantically annotated content that can be used in learning about a general topic, as well as about the user’s needs and interests.

The content structuring and accessing methods are substantially modified if new concepts as *learning objects* (viewed generically as a content object, knowledge object or reusable information object) and *metadata* (object

attributes) are applied. These open new perspectives for the e-learning systems such as personalization, user centred, content reusability in order to compose different curricula or learning path and dynamic content update [2, 3].

The learners have different learning styles, objectives and preferences, which lead to variances of efficiency and effectiveness of the traditional e-learning systems from individual to individual. The learning personalization becomes an advanced stage in the e-learning systems evolution.

The study carried out within the framework of the SinPers project (www.ici.ro/sinpers/) shows that the concept of personalization can be interpreted and implemented in different ways and proves that learning personalization needs new solutions for a multitude of aspects, such as: the identification of the profile, goals and user context; the formalization of knowledge; the description of learner competences and learning objectives; the evaluation of learner’s skill level; feedback return in an adequate manner.

Within the SinPers framework, the personalization issue was solved by adopting innovative solutions in three main domains (figure 1):

1. the modelling of teaching-learning process,
2. learner modelling and
3. digital content modelling.

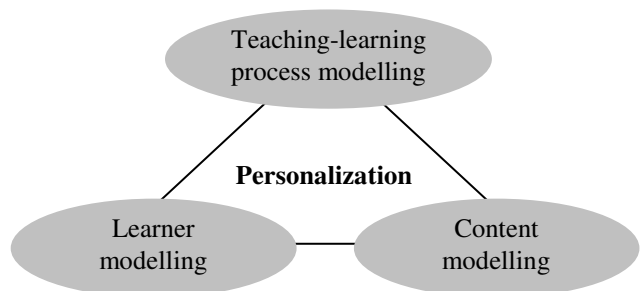


Figure 1. SinPers - personalization pillars

A useful support was the adoption of the IMS standard, which offers a conceptual framework for all three mentioned areas of expertise [4]. This choice was based on a global evaluation of the existing e-learning standards (e.g. SCORM, IEEE, IMS).

II. TEACHING-LEARNING PROCESS MODELLING

The objective of the SinPers project is the development of an e-learning system, centred on the learner and based on advanced technologies. The project aims to surpass several existing limits of the e-learning systems, mainly those referring to the flexibility and adaptability of the training-learning process and to the traditional learning methods, providing new capabilities to the different

Tools for Digital Content Management

Dragos Barbu, Dragos Nicolau, Laura Ciocoiu, Ionut Petre
Research Institute for Informatics, Bucharest, Romanias
ciocoiu@ici.ro

Abstract— The main objectives of this paper are the defining of the principles, the guidelines and the requirements for the development of an integrated collaborative framework for the creation and the management of heterogeneous documents. The paper describes a system with function for messaging, applied for digital library as the assisting method of consulting the users (readers) for acquiring the information and knowledge more effectively and in time, and a Virtual Reference System, that is the platform for virtual library consulting (the readers can get the answers from the manuscripts information experts by the means of email, online questions, viewing FAQ), thus achieve the knowledge accumulation

I. INTRODUCTION

The term “content management” originated in the mid-1990s, and it has several different meanings in today’s marketplace. At its most generic, a content management system is one that stores digital content for search, browsing, access, and retrieval by users in a workgroup or enterprise. [1] The most prevalent types of content management systems are:

- o Digital Asset Management (DAM): systems that manage rich media assets, often including digital audio and video clips, for retrieval and repurposing in media production environments. These systems are sometimes also called Media Asset Management (MAM).
- o Web Content Management (WCM): tools that provide page template design, editorial workflow, and publishing environments specifically for Web sites and other forms of Internet content delivery.
- o Enterprise Content Management (ECM): systems that facilitate management of corporate documents and other types of information for use internally as well as externally with a company’s business partners, customers, regulators, and the general public.

II. THE PROPOSED SYSTEM

The characteristics of proposed system (the framework)can be summarized as follow:

- to include a Content Management System that support multilingualism and provide collaborative tools, e.g. such as forums, chat FAQ; by using these tools the users could more efficiently and easily acquire information and knowledge, resolve their problems and obtains response to their answers; furthermore, by using these tools, the users knowledge and know-how will be collected in the CMS.
- to allow users, by using the CMS, to add their own documents, and to modify or extend the existing ones; the contents could be extended by adding **metadata**, i.e., structured information related to the document. Based on communities; the experiences collected in the last years by several research institute have demonstrated the power of the communities and the need to create platform where the users are protagonist, whose participation becomes key success factor when they create active communities that add, modify and maintain contents collections.

Furthermore the collaboration between community members, by means of forums, newsgroups, chats, FAQ, will transfer the community knowledge inside the platform.

- to allow users to share their files by using P2P (peer-to-peer) technologies interoperable with external tools and services; this means that the frameworks provide a communication mechanism (typically a set APIs) that allows other tools and external service - to invoke it, and, at the same time, the framework can invoke external services and tools. accessible from several platforms such as PCs, PDA and portable phones

- to support online searching and indexing, allowing users to create their own folksonomies on the library documents; (folksonomy is a neologism for define a free and unstructured user defined collaborative categorization using freely chosen keywords)

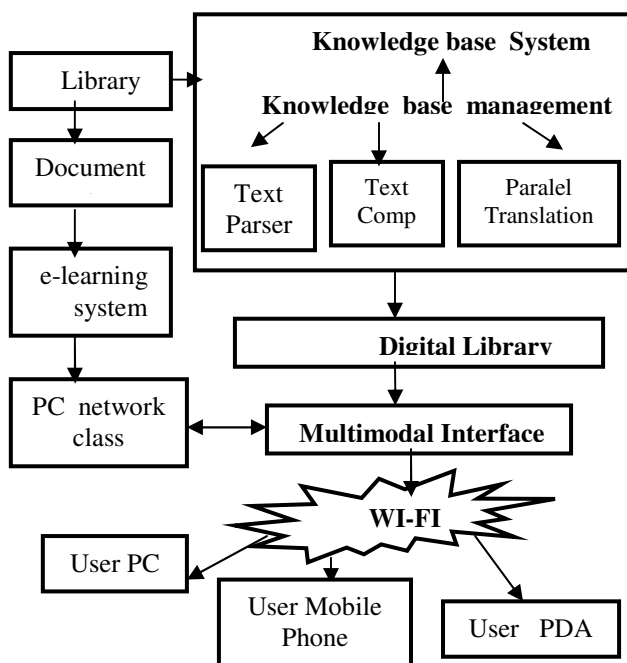


Figure 1 The components of the system proposed

- to adopt of an innovative iDRM (interoperable Digital Right Management) approach, to unobtrusively and efficiently manage IPR/copyright issues while managing the content.

The iDRM approach could be implemented through the AXMEDIS framework, a technology platform developed as result of the AXMEDIS FP6 STREP, that enable content providers and distributors to reduce the costs of cross media production by accelerating and automating the production process. AXMEDIS provides mechanisms for protecting and controlling the use of digital contents, based on MPEG21 and ODRL, and in a scheduled and automated manner and to manage and monitor content’s workflow using the AXMEDIS Workflow Management tool. The AXMEDIS framework provide innovative

Revalorization of National Patrimony through Digital Content Management

Gheorghe Sebestyén-Pál, Bogdan Moscaliuc, Christian Ludovic Nagy

Technical University of Cluj-Napoca, Romania

Gheorghe.Sebestyen@cs.utcluj.ro, Bogdan_Moscaliuc@yahoo.com, Christian.Nagy@yahoo.com

I. INTRODUCTION

Statistics show that today the Internet is the primary source of information not only for the general public but also for researchers in different fields. In this context a document or an information is valuable and it is put in circulation as long as it can be retrieved and reached through the Internet. Mainly for the younger generation the following syntagm is becoming a reality: something exists only if it's on the Internet. This fact put into a disadvantageous position those documents preserved in national archives and classical libraries. Revalorization of these sources of information, mainly of those documents which are part of a national or international patrimony is a major goal, achievable through modern ITC tools.

But, building a digital replica of a classical library through a digitization process (e.g. scanning, OCR, etc.) is not the answer for the problem. It is not enough to transform a document into a digital file. The document must be visible on the Internet, it must be put into circulation and what is even more important, its content or "essence" must be identified and retrieved through specialized search tools. We have to accept that a document is not just an "ASCII string"; it contains notions, concepts, facts that are in a complex relation with each others. A researcher must have adequate tools for semantic search of the digital content.

In the last years an important effort was made in the direction of developing new techniques and algorithms for search, data mining and information retrieval. The most common techniques use key-words search or hierarchical classification of concepts and their related domains. The user must specify a sub-domain, a set of relevant key-words or other classifying or restrictive attributes. Such an approach depends not only on the subjective ability of the user to specify the right key-words or sub-domains, but also on the skills of a librarian to classify and annotate properly the document. Most search tools offered by today's digital libraries make a quest on the metadata associated with a document (e.g. metadata specified by Dublin Core or Marc 21).

Full-text search is a significantly more costly procedure and therefore it is offered as an option just in few cases of digital library management applications. Scanning the full repository of a library generates a significant load on the library server and the answer may be generated in minutes or even in hours. Solution to this problem is twofold: parallelization of the search procedure (e.g. on GRID, cloud of parallel architecture) and implementing more "intelligent" search techniques.

Our research activity in this domain tried to solve some of the issues presented above, in the following way:

- implementation of annotation techniques for the digitized version of old documents, that emphasize the semantic content of these documents

- adoption of an ontology in order to organize content based on concepts and their relations
- implementation of full-text search techniques, based on statistical properties of documents
- reduce the search time through parallelization on GRID infrastructure.

II. PROTOTYPE OF A CONTENT MANAGEMENT SYSTEM FOR ARCHIVES

A. Conceptual view of the proposed system

Figure 1 shows the main components of the proposed digital content management system, intended to enhance search and information retrieval in archives.

The **digitization process** is responsible for transforming the original, mainly paper-based, documents and artifacts into a digital form. Documents are scanned and parts of them are passed through an OCR (Optical Character Recognition) module, in order to obtain a more manageable text format. An important part of the digitization process is the association of meta-data and annotations regarding the content of documents. Part of this job is a traditional **cataloging and classification** operation made by a librarian. There are a number of metadata standards (e.g. Dublin Core, Marc 21, etc.) that allow uniform cataloging and bibliographic data exchange between different library applications. In our approach Dublin Core was used because of its simple structure and widely acceptance in the librarian community. More interesting for archives' revaluation is the annotation process. A librarian or more often a researcher analyzes the documents (images of scanned documents, or text) and makes comments (annotations) concerning specific parts of a document (e.g. names, locations, pictures, etc.).

The annotation module assists the user in associating annotations to the document, through a graphical interface. These annotations are very important mainly in the case of scanned documents that cannot be transformed into a text format through an OCR process (e.g. because of the special character fonts of the document's degradation). In this case the annotation serves for document indexing and later for the information search and retrieval.

In the **annotation process** the user is assisted by an ontology that contains concepts related to the domain of study and relations between them. In this way annotations express the semantics of the analyzed document. A document may be annotated by different persons, so in time, it is enhanced with the multiple views of the readers. In order to keep track of the contributors and their annotations every annotation object preserves identification data about the person who made the adding and the date. So later, a new reader may see all the annotations or decide to filter them based on persons who made the annotation or the date

Intelligent Component for adaptive E-learning Systems

Rodica Potolea, Camelia Vidrighin Bratu, Florin Trif

Technical University of Cluj-Napoca
 {Rodica.Potolea; Camelia.Vidrighin}@cs.utcluj.ro
 Trif.Gelu@dppd.utcluj.ro

Abstract— Most of the existing e-learning systems are based on the quantity of information. They fail to consider an important factor for the success of the system, namely the user. An intelligent adaptive system adjusts the content in a similar manner as face to face education, in order to ensure faster and better performances in the learning process. Moreover, it helps students to develop new, desirable learning abilities. This paper presents the design of a new adaptive elearning system, introducing also the reasons behind the choice of the design. The role and functionalities of each module are described.

I. INTRODUCTION

Education is the driving force in a society. For centuries, face to face instruction has proven to be the best model for conveying knowledge and sustaining the education growth. Even nowadays, this is the only way to ensure the best possible results in this area.

Triggered by the evolution of the technology, especially over the last two decades, other available options have become available, such as online education. Starting from the original goal of knowledge transfer, this paradigm has confirmed to be a valuable instrument in various areas. However, even today, it focuses mainly on content presentation, while lacking the quality of teaching. Therefore, it is mostly used as an alternative instrument, for continuous learning, or as a complementary tool in face to face education.

Current trends in this area focus on the design of elearning systems that contribute to the improvement the user's performance during the learning process. Therefore, the goal is not only to acquire notions, but how to do it in the most appropriate manner for each individual.

This paper presents the features that should be considered when designing an intelligent elearning system, the way they can be combined, and the basic bricks of such a system. We also present a preliminary system and discuss the reasons behind the design obtained.

A brief overview of the most prominent systems which employ intelligent components to model the user's type and to adapt accordingly is presented in the following. In [6, 11], the pedagogical diagnosis is performed by hybrid agents; SQL Tutor+ [16], LearnOOP [15], IOLS [5] employ intelligent agents to determine the user's profile; AdaptaTIVE [10] makes use of an intelligent component to automatically categorize content; SIETTE [3], a testing system, adapts the queries to the user's level; Betty's Brain [1, 17] is based on the development of self-regulation abilities.

The rest of the paper is organized as follows: section 2 presents learning styles and their importance in the learning process. Types of elearning systems, their advantages and disadvantages are discussed in section 3. Section 4 presents the model of an adaptive elearning system, and section 5 presents our design of the modules

presented in section 4. The conclusions and future work directions are discussed in section 6.

II. LEARNING STYLES

Most of the online training systems are based on curricula segmentation, situation in which students must go through a predefined structure. It is widely acknowledged that the student should be involved actively in the online learning process and that elearning systems should sustain the student's control and organization of information [7]. Thus, the online systems should constrain the user less and (occasionally) allow him/her to decide which path to follow.

In face to face education, the teacher conducts the education process by analyzing the students' level of understanding, their feedback, reactions, and also by his/her (the teacher's) feeling. Therefore, the education is specialized according to the quality of the students in the class. Traditional elearning systems are mostly electronic versions of textbooks, with some communication enhancements. Hence, most students, regardless of their background, motivation, learning abilities and learning style, have to follow the same path, and in most situations, at the same pace.

In order to deal with the situation of differential learning, an adaptation process should be considered, according to the student's learning style. Therefore, some degree of freedom in both the quantity and quality of the material to follow should be allowed. Moreover, the level of difficulty should be elective as well.

With the Inventory of Learning Styles, Vermunt [13] aimed to integrate different learning processes, some of which are thought to be relatively stable (mental learning models and learning orientations) and some of which are contextually determined (choice between regulatory and processing strategies). This model applies to the thinking and learning activities performed by university students. It is experientially grounded in interviews with students and seeks to integrate cognitive, affective, meta-cognitive and co-native processes. It is dependent on context, so a learning style is the interplay between personal and contextual influences. The accent moves from an individual differences approach to the whole teaching-learning environment.

The inventory consists of two parts, Part A and Part B. Part A, called Study Activities, includes questions on two domains, processing strategies and regulation strategies. Part B, called Study Motives and Views on Studying includes two parts: B1, study motives, referring to learning orientations and B2, study views, which addresses the mental models of learning. This model proposed by Vermunt contains four components: meaning-directed, reproduction-directed, undirected and application-directed learning styles. Each of the four components is further divided into five subscales, containing between four and six items.

Managing Knowledge and Quality in Higher Education Study Programmes

Claudiu Vasile Kifor, Lucian Lobont, Constantin Oprean

Hermann Oberth Faculty of Engineering

Lucian Blaga University of Sibiu

Sibiu, Romania

e-mail: claudiu.kifor@ulbsibiu.ro

lucian.lobont@ulbsibiu.ro

rector@ulbsibiu.ro

Abstract - In today's world, all the economy domains, functions and elements are constrained by qualitative changes given by the circumstances of the information society and supported by knowledge, learning and innovation. In order that the society trusts the universities abilities to satisfy its expectations, the universities must take full responsibility for the quality of their own actions. Due to the fact that the innovative potential of Information and Communication Technologies (ICT) in the quality management (QM) area has been little exploited, we have developed and we intend to implement, via a pilot project, ICT tools based on knowledge and learning. These ICT tools are to provide the basis of the QM system design, implementation and continuous improvement in the higher education institutions. The paper presents the achievements of the e- EdU - Quality project, whose objective is to improve the academic processes through an integration of ICT and quality management tools. The concrete result of this project was a platform implemented at the Master Programme in Quality Management from Lucian Blaga University of Sibiu.

Keywords-quality; knowledge; higher education;

I. INTRODUCTION

A Romanian Education Ministry Regulation says that: "each higher education institute from Romania (...) is the main responsible for the quality of educational services and also for quality assurance, (...) having the obligation to apply its own quality assurance system starting with the 2005/2006 academic year" [1]. On September 2003, the ministers of education from the EU countries met in Berlin and decided that all member and candidate countries should define and implement a national quality assurance system in higher education by 2005. In order to be involved in European projects for research and mobility, the universities should reach a consensus on their structure and offer content, on standards, criteria and procedures for results evaluation. This is also valid in order to get international recognition of diplomas. At the European level, we can underline different stages for quality assurance and QM in education [2]:

- Sorbonne declaration (1998) referring to the European Space for Higher Education;
- Bologna declaration (1999) is a pledge by 29 countries to reform the structures of their higher education systems in a convergent way until 2010 by promoting European cooperation for quality assurance;
- Lisbon convention (2000), pinpoint the idea of a competitive economic society based on knowledge;
- Salamanca convention (2001);

- Berlin declaration (2003), where has been established that "the main responsibility for quality assurance in higher education belongs to each institution";
- Bergen convention (2005) made a set of recommendations for guides, standards and procedures, national and international framework for qualifications, minimum number of ECTS, life long learning.

At the European level the quality assurance agencies, network and associations such as European Association for Quality Assurance in Higher Education (ENQA), the Nordic Quality Assurance Network in Higher Education (NOQA), the Network Agencies for Higher Education Quality Assurance from Central and Oriental Europe, the Network D-A-CH, continuously develop tools and methods for quality evaluation, assurance and improvement.

UNESCO and OECD are also involved to elaborate and adopt some "guidelines" regarding the quality of services offered by trans-national education institutions [3]. The interests for quality and e-learning technologies reside also from the national initiatives like e-Fit Austria. Some of the main objectives of this initiative are: easy access to innovative services, high quality content for education, science, training and culture, all these using the information technology to create a better and more efficient service for the educational system.

At the national level even the quality assurance activities were previously developed, the focus on higher education quality assurance is given by the creation of the Agency for Quality Assurance in Higher Education (ARACIS) [4] a member of ENQA from June 2009 [5]. The relations between the European education framework and national education frameworks are based on transparency, visibility and compatibility, therefore any higher education institution is responsible to develop a culture for quality. This means politics, techniques and practices consequently applied and documented to obtain results in concordance with the proposed objectives.

II. DEVELOPMENT OF E-EDU-QUALITY

Starting from the existing realities and challenges at national and international level, e - EdU - Quality project [6] has the following general objective: improvement of academic QM and of the academic education's quality through the implementation of informatics systems based on learning and knowledge.

The research aims to develop and implement, through a pilot project, informatics tools for knowledge and learning meant to represent the basis for the QM system performance and continuous improvement within the higher education institutions. This initiative represents a

Applying Rule-based Knowledge Acquisition on Historical Documents

Ioan Salomie¹, Cristina Pop¹, Tudor Cioara¹, Ionut Anghel¹, Marius Grindean¹,
Ioan Dragan², Livia Ardelean², Ioan Carja³

¹Department of Computer Science, Technical University of Cluj-Napoca, Romania

²Cluj County National Archives, Romania

³Faculty of History and Philosophy, Babes-Bolyai University, Cluj-Napoca, Romania

Contact: Ioan.Salomie@cs.utcluj.ro

Abstract— This paper presents a rule-based knowledge acquisition method for extracting and semantically annotating the relevant information from text documents. We use a domain ontology as knowledge base for persisting the extracted information. The rule-based knowledge acquisition method relies on a set of lexical annotation rules and on a set of semantic annotation rules. The lexical annotation rules are used in information extraction, while the semantic annotation rules are used for mapping the extracted information to the elements of the domain ontology. The rule-based knowledge acquisition method was experimented on a set of Romanian archival documents providing good results that can be used in knowledge retrieval.

Keywords - knowledge acquisition; rule; lexical annotation; semantic annotation; ontology.

I. INTRODUCTION

For us, humans, everything that is displayed on the screen has a meaning, but for computers plain text and images are useless unless we provide additional information. An efficient and popular solution is annotation. This means that the primary information is associated to a semantic description using tags allowing systems to analyze the information and perform reasoning. Transforming the raw information in computer-processable data is a task called knowledge acquisition [1], an important task for knowledge based systems because a lot of training data is required to infer the right decision.

Our research target is a rule based knowledge acquisition method for a natural language processing system. Two major problems were identified: metadata definition and mapping the relevant content of the documents to the content descriptors. The metadata is used in the annotation process of the input documents. The extracted information must be mapped to the knowledge base which contains content descriptors emphasizing the semantics of the processed data.

We propose a solution for these problems by using a grammar of lexical annotation rules and a predefined dictionary. The lexical annotation rule contains two parts, one defining the context (when to apply the rule) and one defining how to treat the identified context. In case of text documents, the first part of the rule is a pattern that matches a part of text and the second part of the rule is a tag that offers a piece of information about that part of text. The mechanism works like a classifier: whenever a part of the text matches the rule's pattern, the associated annotation is assigned.

However, these lexical annotations are not enough because semantics is not precisely preserved. In order to overcome this problem, semantic annotation based on a domain ontology is considered. Associating ontology concepts and relations to raw text offers a means of

defining semantics. Thus, it is necessary to provide a method in order to map text to semantic information.

Our proposed solution to this problem uses a set of semantic annotation rules. The purpose of these rules is to enrich the domain ontology with new instances, relationships and other ontology concepts. For each lexical annotation rule we need to define/provide a semantic annotation rule. Semantic annotation rules have two parts also, one for context specification and one for associated actions that will be performed on the ontology. The first part identifies the tag associated with the appropriate lexical annotation rule and the second part of the rule contains ontology enrichment actions. An ontology is a machine-processable abstract model which includes concepts, instances and their associated relations. The ontology captures knowledge independently of its use and in a manner that makes it possible to share the mentioned knowledge.

Thus, two types of rules are involved in the knowledge acquisition process. First lexical information is extracted from raw data by means of lexical annotation rules. In this case lexical information refers to a part of the text and the associated mark-up. Next, semantic annotation rules map lexical information to ontology concepts. In order to obtain a critical mass for our knowledge base, a consistent set of data acquisition rules is necessary.

The solution described in this paper is part of the ArhiNet research project [2]. The main objective of the ArhiNet project is the study and development of an integrated system for generating and processing semantically enhanced e-content from document archives. Using a domain ontology, semantically annotated documents and raw content, relevant information can be retrieved using a natural language query system. The proposed system is able to perform knowledge acquisition from primary historical sources available in various formats. The information is extracted and translated into knowledge used to populate a domain specific ontology. The provided knowledge will be further used by reasoning algorithms in order to provide new information. To make use of the obtained knowledge, ontology-guided query algorithms were developed.

The rest of the paper is organized as follows. Section II contains related work in this study area. Section III presents our conceptual solution of creating data acquisition rules. Section IV contains implementation details of the rule creation workflow. Section V contains a case study that illustrates the rule creation workflow in the context of historical archives. The paper ends with our conclusions and future work proposals.

II. RELATED WORK

Rule-based knowledge acquisition has been studied and applied in several systems and methodologies which will be described in what follows.

Applying Rule-based Knowledge Acquisition on Historical Documents

Ioan Salomie¹, Cristina Pop¹, Tudor Cioara¹, Ionut Anghel¹, Marius Grindean¹,
Ioan Dragan², Livia Ardelean², Ioan Carja³

¹Department of Computer Science, Technical University of Cluj-Napoca, Romania

²Cluj County National Archives, Romania

³Faculty of History and Philosophy, Babes-Bolyai University, Cluj-Napoca, Romania

Contact: Ioan.Salomie@cs.utcluj.ro

Abstract— This paper presents a rule-based knowledge acquisition method for extracting and semantically annotating the relevant information from text documents. We use a domain ontology as knowledge base for persisting the extracted information. The rule-based knowledge acquisition method relies on a set of lexical annotation rules and on a set of semantic annotation rules. The lexical annotation rules are used in information extraction, while the semantic annotation rules are used for mapping the extracted information to the elements of the domain ontology. The rule-based knowledge acquisition method was experimented on a set of Romanian archival documents providing good results that can be used in knowledge retrieval.

Keywords - knowledge acquisition; rule; lexical annotation; semantic annotation; ontology.

I. INTRODUCTION

For us, humans, everything that is displayed on the screen has a meaning, but for computers plain text and images are useless unless we provide additional information. An efficient and popular solution is annotation. This means that the primary information is associated to a semantic description using tags allowing systems to analyze the information and perform reasoning. Transforming the raw information in computer-processable data is a task called knowledge acquisition [1], an important task for knowledge based systems because a lot of training data is required to infer the right decision.

Our research target is a rule based knowledge acquisition method for a natural language processing system. Two major problems were identified: metadata definition and mapping the relevant content of the documents to the content descriptors. The metadata is used in the annotation process of the input documents. The extracted information must be mapped to the knowledge base which contains content descriptors emphasizing the semantics of the processed data.

We propose a solution for these problems by using a grammar of lexical annotation rules and a predefined dictionary. The lexical annotation rule contains two parts, one defining the context (when to apply the rule) and one defining how to treat the identified context. In case of text documents, the first part of the rule is a pattern that matches a part of text and the second part of the rule is a tag that offers a piece of information about that part of text. The mechanism works like a classifier: whenever a part of the text matches the rule's pattern, the associated annotation is assigned.

However, these lexical annotations are not enough because semantics is not precisely preserved. In order to overcome this problem, semantic annotation based on a domain ontology is considered. Associating ontology concepts and relations to raw text offers a means of

defining semantics. Thus, it is necessary to provide a method in order to map text to semantic information.

Our proposed solution to this problem uses a set of semantic annotation rules. The purpose of these rules is to enrich the domain ontology with new instances, relationships and other ontology concepts. For each lexical annotation rule we need to define/provide a semantic annotation rule. Semantic annotation rules have two parts also, one for context specification and one for associated actions that will be performed on the ontology. The first part identifies the tag associated with the appropriate lexical annotation rule and the second part of the rule contains ontology enrichment actions. An ontology is a machine-processable abstract model which includes concepts, instances and their associated relations. The ontology captures knowledge independently of its use and in a manner that makes it possible to share the mentioned knowledge.

Thus, two types of rules are involved in the knowledge acquisition process. First lexical information is extracted from raw data by means of lexical annotation rules. In this case lexical information refers to a part of the text and the associated mark-up. Next, semantic annotation rules map lexical information to ontology concepts. In order to obtain a critical mass for our knowledge base, a consistent set of data acquisition rules is necessary.

The solution described in this paper is part of the ArhiNet research project [2]. The main objective of the ArhiNet project is the study and development of an integrated system for generating and processing semantically enhanced e-content from document archives. Using a domain ontology, semantically annotated documents and raw content, relevant information can be retrieved using a natural language query system. The proposed system is able to perform knowledge acquisition from primary historical sources available in various formats. The information is extracted and translated into knowledge used to populate a domain specific ontology. The provided knowledge will be further used by reasoning algorithms in order to provide new information. To make use of the obtained knowledge, ontology-guided query algorithms were developed.

The rest of the paper is organized as follows. Section II contains related work in this study area. Section III presents our conceptual solution of creating data acquisition rules. Section IV contains implementation details of the rule creation workflow. Section V contains a case study that illustrates the rule creation workflow in the context of historical archives. The paper ends with our conclusions and future work proposals.

II. RELATED WORK

Rule-based knowledge acquisition has been studied and applied in several systems and methodologies which will be described in what follows.

Information Extraction from Romanian Archival Documents

Mihaela Dinsoreanu¹, Tudor Cioara¹, Ionut Anghel¹, Laura Dinsorean¹, Marius Georgiu¹,
Ioan Carja², Ioan Dragan³, Livia Ardelean³

¹Department of Computer Science, Technical University of Cluj-Napoca, Romania

²Faculty of History and Philosophy, Babes-Bolyai University, Cluj-Napoca, Romania

³Cluj County National Archives, Romania

Contact: *Mihaela.Dinsoreanu@cs.utcluj.ro*

Abstract — This paper presents an information extraction method which processes the content of Romanian text documents to create a domain ontology used for knowledge retrieval. The method uses a natural language processing tool and a set of rules to lexically and semantically annotate the relevant information from text. The natural language processing tool was adapted to process Romanian natural language. The set of rules is used to extract concepts, instances and relations from the text and to automatically enrich the domain ontology. We tested and validated the information extraction method on a set of archival documents containing historical facts about the medieval Transylvania.

Keywords – *information extraction; semantic annotation; ontology enrichment*

I. INTRODUCTION

Archives represent a rich and valuable source of historical information containing data about events, important personalities and relevant moments in history. Because of the importance and age of the documents only few people have access to them thus finding a way to store them in digital format would protect the original documents and simplify the access and search methods. The archive could be accessed at any time, by any user and the document search time on a certain topic would be reduced. With large document databases, such as archives, the real challenge is finding relevant information regarding a certain topic. The documents are written in natural language and although it is easy for humans to understand them, computers can only interpret the structure of the used words. A possible solution to the problem is proposed by the semantic Web approach [6]. The enrichment of documents with semantic information allows computers to extract meaningful information. The semantic enrichment of documents improves information management, exchange and search.

ArhiNet [7] is an integrated system for generating and processing semantically enhanced archival documents. The system builds a domain ontology based on a set of initial knowledge, and the information extracted from the archival documents. By analyzing the available corpus we developed a core ontology capturing the most general concepts encountered in the documents content. The ontology is further enriched using reasoning techniques and is used to semantically annotate the input documents. The semantic annotations and the domain ontology provide the necessary support to execute semantic queries, which provide more relevant results than the classical searching approach.

In this paper we describe one method we investigated for information extraction from Romanian archival documents in the ArhiNet system. The input documents are lexically annotated using natural language processing methods [1] and then semantically annotated based on a set of rules that identify common language patterns. Based on

the annotations, instances, concepts and relations are extracted for automatically enriching the domain ontology [2].

The rest of the paper is organized as follows. In Section II we introduce related work. In Section III we present the architecture of the information extraction and data acquisition module. The system's initial ontology is described in Section IV. The raw input documents are lexically annotated using a GATE pipeline which is described in Section V. Section VI presents the JAPE rules used to semantically annotate the archival documents. In Section VII we present a case study that illustrates the system functionality. In Section VIII we present some general results and in Section IX we draw a conclusion and give some future work proposals.

II. RELATED WORK

ANNIE (A Nearly-New Information Extraction System) [5] is an information extraction system composed of a set of GATE (General Architecture for Text Engineering) components: tokenizer, gazetteer, sentence splitter, part of speech tagger and coreference tagger. The components form a pipeline that processes a text document. The input text is segmented into sentences and tokens such as words, numbers, white characters etc. Every word is searched in a set of gazetteer lists to find its category and in a lexicon to find its part of speech.

Although some GATE components have also been developed, as a prototype, for the Romanian language, ANNIE is not yet able to extract information from Romanian documents. The ground differences between the English and Romanian grammars make it difficult to use the same information extraction pipeline. In this paper we have developed a solution that uses a part of the ANNIE pipeline (tokenizer, sentence splitter, part of speech tagger) in order to lexically annotate the input documents and prepare them for further processing.

Text2Onto [4] is a framework and also application for ontology learning from text documents. The framework uses natural language processing techniques to linguistically analyze the input text, implemented as a pipeline of GATE components: tokenizer, sentence splitter, part of speech tagger and lemmatizer/ morphological analyzer. Based on the linguistic annotations, the text is semantically annotated using a set of JAPE rules which identify common language patterns. The final processing step consists of applying a set of algorithms and combiners in order to extract concepts (classes), instances, concept inheritance, concept instantiation and relations which are further used to enrich the domain ontology.

Text2Onto is only available for processing English documents and it is currently in development for Spanish and German. Because of the very good results proven in automatic ontology learning we have chosen to use this framework in the solution described in this paper. In order to adapt it to the Romanian archival documents we had to implement two of the GATE components (part of speech

Hierarchical Data Models for the ArhiNet Ontology Representation and Processing

Mihaela Dinsoreanu¹, Ioan Salomie¹, Cornel Rat¹, Sorin Suciu¹, Tudor Cioara¹, Ionut Anghel¹,
Ioan Dragan², Livia Ardelean², Ioan Carja³

¹*Department of Computer Science, Technical University of Cluj-Napoca, Romania*

²*Cluj County National Archives, Romania*

³*Faculty of History and Philosophy, Babes-Bolyai University, Cluj-Napoca, Romania*

Contact: *Mihaela.Dinsoreanu@cs.utcluj.ro*

Abstract—In this paper we present a hierarchical data model for representing and processing ontologies. By using this model, hierarchical relationships like class or property inheritance, frequently encountered in the internal structure of an ontology, are efficiently handled. The proposed hierarchical data model defines a generic representation using hierarchies which can be persisted in a general purpose records structure. We also propose two algorithms for creating and persisting the hierarchical relationships encountered in ontologies into the persistent hierarchical data model. The performance of hierarchical queries was tested on an archival domain ontology which was persisted using the proposed hierarchical data model.

Keywords—ontology; hierarchical data model; hierarchical relationship; hierarchical query; archival domain

I. INTRODUCTION

The Semantic Web aims at enhancing the resources available on the Web with machine-understandable descriptions and means of interchange. A solution to achieving systems interoperability is provided by the use of ontologies as common vocabulary and domain model.

The Web Ontology Language (OWL) [10] is a World Wide Web Consortium (W3C) [11] Recommendation and therefore it has become a widely used data format for representing ontologies. In time, large domain ontologies have been developed, leading to the need of finding efficient ways of storing and querying OWL ontologies.

One of the main issues with large ontologies is the way of accessing data. In the case of file stored OWL ontologies, the access to a given ontology element is a linear function of the file size, therefore the file-based storage approach is not a feasible solution. To overcome the access time problem two main solutions have been developed: an in memory ontology representation on one hand and database ontology representations on the other hand. The in memory solutions, feature two main drawbacks: (i) for large ontologies the memory may become an issue and (ii) the re-parsing overhead which occurs each time the ontology is updated and thus re-loaded into memory. Mapping the ontology into a database has the advantages of (i) persisting the ontology structure to a model that provides fast access times for accessing its elements and (ii) the possibility of using the database associated query languages for retrieving various data from the ontology. The systems of mapping the ontologies to databases take various approaches. Some of the relevant OWL to relational approaches [6] [7] create database tables for each of the ontology classes and link the tables between them by using foreign key constraints. The table links are based on the appropriate ontology properties (data type and object) affecting the given class. Database metadata and schema information are then used for performing queries on the resulted mapping model. This approach has the advantage

of creating a mapping model closer to the structure of the OWL class structure, but queries have to be dynamically generated and the advantage of compiled stored procedures is lost. Also, the usage of several database tools such as the Business Intelligence tools for performing data mining, becomes complicated.

Given the high number of hierarchical relations present in ontologies and their frequent use in ontology related operations, hierarchical data models (HDMs) provide means for a natural and efficient way of representing the ontology hierarchical relations.

In this paper we propose a hierarchical data model for efficiently handling hierarchical relationships found both in the internal ontology structure (e.g. class inheritance, property inheritance) which addresses the main drawbacks identified in other approaches. Our hierarchical data model defines a generic representation using hierarchies that can be persisted in a general purpose records structure. We also present an efficient solution to mapping ontologies from OWL format to our proposed data model. In our HDM, the hierarchical ontology relationships are described using trees, while the ontology elements are represented using relational data entities. The HDM allows for multiple ontologies to be represented together. Our general-purpose structure of the persisted model allows its contained data to be used by data analysis tools.

The proposed solution was tested by using the ontology of the ArhiNet research project [9]. The project deals with heterogeneous, multilingual archivist documents spanning across several centuries of the Transylvania region history. ArhiNet has as the end goal the creation of a system that would allow users to efficiently query and process the archivist documents. Reasoning and learning algorithms are used in order to extract further knowledge, improve/enhance the ontology and classify the corpus documents. The process involves the organization and processing of the knowledge found in the documents using a storage environment that allows scalable information querying. Scalability of knowledge processing mechanisms is a necessity for the ArhiNet project because it requires large volumes of ontology data to be handled.

A. ArhiNet System - Knowledge Acquisition Overview

The ArhiNet knowledge acquisition workflow (KAW) (see Figure 1) incorporates our proposed HDM for storing the domain ontology. The objective of the ArhiNet Knowledge Acquisition Workflow (KAW) is to create a digital repository of semantically annotated archival documents that can be further used for machine reasoning and learning. The ArhiNet KAW uses as corpus the of raw documents about the medieval history of Transylvania available at the Cluj

Distributed, Agent-based, Semantic Reasoning Engine for Ontologies

A Case-study in the Archival Domain

Ioan Salomie¹, Cornel Rat¹, Sorin Suci¹, Cristina Pop¹, Viorica Chifu¹,
Ioan Dragan², Livia Ardelean², Ioan Carja³

¹*Department of Computer Science, Technical University of Cluj-Napoca, Romania*

²*Cluj County National Archives, Romania*

³*Faculty of History and Philosophy, Babes-Bolyai University, Cluj-Napoca, Romania*

Contact: Ioan.Salomie@cs.utcluj.ro

Abstract—In this paper we present a distributed architecture for an agent-based reasoner. The reasoner uses ontologies for inferring semantic knowledge and for providing relevant results to user queries. We propose a data model for decoupling the reasoner from the query language used for interrogation, thus enabling multilingual query processing. We use a persistent hierarchical model for efficiently processing the ontology and for delegating the task of handling distributed queries to a database management system. The agent-based reasoner relies on a divide et impera technique for breaking the user queries into smaller subtasks which are assigned to three types of distributed agents for further processing. We have tested the agent-based reasoner within the ArhiNet system for providing the relevant information available in archival documents about the history of Medieval Transylvania.

Keywords-ontology; agent-based reasoner; hierarchical model; distributed agent; archival documents

I. INTRODUCTION

Modern systems have to deal with large amounts of data on various subjects. Some of the data is shared on the Internet, some is available only in printed format, but most of it is stored in a way which is not suitable for machine processing. The problem resides in the fact that data is represented in natural language, which makes it easy for humans to understand it but almost impossible for computers to make sense of it. In order to deal with the problem, information extraction techniques have been developed and used on digital data repositories in order to obtain and structure information in a computer comprehensible manner.

Ontologies constitute a resource aiding computers in understanding and processing natural language data by providing a shared understanding of a specific domain. The information encoded within ontologies is modeled in terms of concepts and relationships between them, instances of the concepts represented as individuals and restrictions. Machines use reasoning engines working on top of ontologies for analyzing natural language content. In this paper we propose a distributed, agent based reasoning engine working on top of an ontology, which can leverage existing technologies and reasoning engines in order to find relevant results to user queries and which can enrich existing ontologies with new concepts based on the information already present in the initial ontology.

The rest of the paper is organized as follows. Section 2 provides an overview regarding the related work. Section 3 introduces our data model used for decoupling the reasoner from the query language and for representing the query as a structure that can easily be processed in a parallel and distributed manner. Section 4 describes the reasoner and introduces the agent types used for processing user queries. Section 5 provides the case study used for validating our

reasoner. Section 6 presents our conclusions and introduces future work areas.

II. RELATED WORK

Many reasoning engines have been developed to work on top of ontologies in order to answer user queries about existing data or for extracting new knowledge from the existing information. A first order logic approach to the problem is presented in [1] [2]. In [1] the authors propose a method of transforming ontology concepts represented using the SUO-KIF language into standard first order logic and propose a set of optimizations which allow FOL reasoners to handle large ontologies. In [2] the author presents a first order logic approach to specifying and querying relational, hierarchical and network databases and presents the advantages of the approach. We, on the other hand, take a different approach and transform first order logic queries into database queries.

In [3] the authors present a method of reasoning on large ontologies represented in relational databases. They provide a DL-Lite reasoner that translates the user queries into SQL queries and thus leverages the power of the relational database management system (RDBMS).

An important fact that has to be considered while developing a reasoning engine is represented by the level of certitude of the data that has to be analyzed. From the certitude perspective, data can be split into certain and uncertain data. A reasoning engine dealing with uncertain information modeled on Bayesian Networks is presented in [4]. The proposed solution uses ontologies for representing the knowledge base. The authors also propose an algorithm for constructing conditional probability tables by starting from an ontology and a set of guidelines. We use data mining algorithms for extracting knowledge from uncertain data and reasoning with it.

Query evaluation and optimization represents another problem in itself while creating reasoning engines. In [5] the authors present a dynamic programming method for optimizing the query execution plans and thus speeding up the evaluation of the query.

A federated approach to reasoning on modular ontologies is presented in [6]. The solution proposes a local reasoner for each of the ontology modules and propose a parallel, asynchronous reasoning algorithm working on Package-based Description Logic (P-DL).

In [7] the authors present a framework for splitting and processing large amounts of data on distributed systems. The solution works on data sets organized as key-value pairs used by the two user defined functions, map and reduce, which partition the data and process and merge the results. We propose a tree structure for modeling the work and the