CONTENTS

Automat. Comput. Appl. Math.
Volume 19 (2010), Number 3

COMPUTERS IN MEDICAL DIAGNOSIS

DIGITAL LIBRARIES, E-CONTENT MANGEMENT AND E-LEARNING

Liver Fibrosis detection by the means of texture analysis. Limitations and further development directions

Cristian Vicas[1], Monica Lupsor[2], Mihai Socaciu[2], Radu Badea[2], Sergiu Nedevschi[1]
[1] Faculty of Automation and Computer Science, Technical University of Cluj-Napoca ROMANIA
[2] 3rd Medical Clinic, University of Medicine and Pharmacy, Cluj-Napoca ROMANIA
cristian.vicas@cs.utcluj.ro

Abstract — **Texture analysis is viewed as a method to enhance the diagnosis power of classical B-mode ultrasound image. Present study aims to evaluate the dependence between the human expert and the performance of such a texture analysis system in predicting the cirrhosis in chronic hepatitis C patients. 125 consecutive chronic hepatitis C patients were included in this study. All the patients had positive HCV-RNA in serum and had undergone percutaneous liver biopsy for disease staging using Metavir score. Ultrasound images were acquired from each patient and 4 experts established regions of interest. Textural analysis software generated 234 features from each region of interest. Relevant textural features were identified and a classification schema was evaluated. Texture analysis can discriminate between F0 and F4 fibrosis stages (AUROC=0.64). The performance of this approach depends highly on the human expert that establishes the regions of interest ($p<0.05$). The relevant textural features were identified and it was shown that the detection performance didn't depend on the particular feature selection ($p=0.8$). In classical form met in literature non invasive diagnosis through texture analysis has limited utility in clinical practice because of the user variability introduced by the expert who establishes the regions of interest.**

Keywords-non invasive diagnosis; texture analysis; user variability; liver fibrosis

I. INTRODUCTION

Non invasive detection and staging of liver fibrosis have received more and more attention in scientific literature. One approach involves simple B-mode ultrasound in conjunction with textural analysis. The main assumption of the textural analysis approach is that fibrosis alterations at liver lobule level can induce significant changes in the speckle pattern of the ultrasound image[1]. Even if these alterations are not visible with naked eye, a texture analysis system can detect and learn these alterations. Textural analysis is viewed as a method to enhance the diagnosis power of B-mode ultrasound by providing the physician with new information. This data can be otherwise inferred only by invasive methods.

The methodology presented in most of the papers[1-5] approaching textural analysis on B-mode ultrasound follows four general steps. First, a physician acquires a liver ultrasound image. Then, on the ultrasound image, another physician (or the same) establishes a rectangular region of interest (ROI). In the third step several textural algorithms produce a feature vector. This vector is labeled according to biopsy findings. The fourth step implies the training of a classification schema. The resulting classifier can be used to predict fibrosis stages to unknown ultrasound images. In the first two steps there is a human expert that introduces an operator dependent variability.

This paper addresses the user variability introduced by the second step, the establishment of the ROI. To our knowledge this problem has not been addressed before. We included almost all the textural algorithms proposed in the literature as means of detecting liver fibrosis stages.

Present study aims to evaluate the dependence between the human expert and the performance of the texture analysis system in predicting cirrhosis in chronic hepatitis C patients.

II. MATERIAL AND METHODS

A. Patients

This study was approved by the local Ethical Committee of the University of Medicine and Pharmacy Cluj-Napoca. The patients provided written informed consent before the beginning of the study, in accordance to the principles of the Declaration of Helsinki (revision of Edinburgh, 2000). We prospectively included in this study 125 patients with hepatitis C infection having fibrosis stage 0 or 4 according to Metavir scoring system. The fibrosis stages were determined by liver biopsy. This lot was selected from 1200 patients, prospectively examined in 3rd Medical Clinic, Cluj-Napoca, Romania, between May 2007 and August 2009. All patients had positive HCV-RNA and underwent percutaneous liver biopsy (LB), in order to stage and grade their condition.

The exclusion criteria were: presence of ascites at clinical or ultrasound examination, co-infection with HBV and/or HIV, other active infectious diseases, and pregnancy.

Alongside the epidemiological data, certain biological parameters were determined on a blood sample taken 12 hours after overnight fasting: alanine aminotransferase (ALT), aspartate aminotransferase (AST), gama-glutamyl transferase (GGT), total cholesterol, triglycerides, total bilirubin and glycemia (Konelab 20i – Thermo Electron Corp., Finland).

B. Histopathological analysis

A liver biopsy specimen was prelevated using the TruCut technique with an 1.8 mm (14G) diameter automatic needle device - Biopty Gun (Bard GMBH, Karlsruhe, Germany). The LB specimens were fixed in formalin and embedded in paraffin. The slides were evaluated by a single expert pathologist unaware of the clinical data. Only biopsy specimens with more than 6 intact portal tracts were eligible for evaluation . The liver fibrosis and necroinflammatory activity were evaluated semi quantitatively according to the Metavir scoring system[6].

Fibrosis was staged on a 0-4 scale as follows: F0 – no fibrosis; F1 – portal fibrosis without septa; F2 - portal fibrosis and few septa; F3 – numerous septa without cirrhosis; F4 –

Relevant feature selection and automatic recognition of the inflammatory bowel diseases and colon cancer from B-mode and contrast-enhanced ultrasound images

D. Mitrea[1], P. Mitrea[1]
[1]Computer-Science Department
Technical University of Cluj-Napoca,
Cluj-Napoca, Romania
Delia.Mitrea@cs.utcluj.ro, Paulina.Mitrea@cs.utcluj.ro

M. Socaciu[2], L. Ciobanu[2], R. Badea[2], A. Golea[2], C. Hagiu[2], A. Seiceanu[2]
[2]Department of Ultrasonography,
3rd Medical Clinic,
Cluj-Napoca, Romania
socacium@yahoo.com, rbadea2003@yahoo.com

Abstract— **The inflammatory bowel diseases (IBD) are severe, chronic and recurring disorders. The most reliable methods for the diagnosis of these diseases are invasive (endoscopy, colonoscopy, histopathology) or irradiating (CT), dangerous for the patient. We develop computerized methods for the assessment of the bowel inflammation level based on information obtained from ultrasound images. In this work, we aim to find the relevant features which are appropriate for the characterization and recognition of the inflammatory bowel diseases and colon cancer. Several feature selection methods are compared for this purpose. Image enhancement techniques, textural parameters and time intensity curves (TIC) based parameters are taken into consideration. The relevant textural and TIC parameters are considered individually and in combination in order to assess the classification accuracy. B-mode and contrast-enhanced ultrasound images (CEUS), belonging to biopsied patients, are used in our study. The considered features proved to be efficient, leading to satisfying results.**

Keywords – inflammatory bowel diseases; texture; Time-Intensity Curves (TIC); relevant features; automatic diagnosis

I. INTRODUCTION

The inflammatory bowel diseases (IBD) are a group of disorders that often mark the population of the developed countries. Their evolution is frequently chronic, with activation peaks and remission periods, elements that are conditioned by the rapidity of the diagnosis, by the follow-up efficiency and by the therapeutic means. There are several clinical, laboratory and para-clinical parameters used to assess the activity phase. Most used clinical scores are the Crohn's Disease Activity Index (CDAI) and Truelove Witts. Together with laboratory parameters, 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, while the other structures are inaccessible. Computer tomography and magnetic resonance imaging are elective imagistic methods, but are less accessible and quite expensive. Ultrasonography has similar potential in diagnosis, but with advantages like: non-invasivity, reduced cost and the possibility of repeatability. Numerous literature studies have proven their role for the examination of the digestive tube pathology [1], [2], [3]. Our aim is to develop new methods of activity assessment in inflammatory bowel diseases, based on utrasonography examination, combined with some modern techniques like vascular contrast enhancement and computer-aided analysis of images. First, some image enhancement techniques are implemented in order to emphasize the specific visual features for each kind of disease, within B-mode ultrasound images. Also, texture is considered, as an important visual feature, able to provide subtle information concerning the structure of the internal organ tissues. We develop texture-based methods in order to emphasize the features that characterize each inflammatory bowel disease and the digestive tumors. Both relevant feature selection and automatic recognition are involved in our research. Contrast-enhanced ultrasound (CEUS) provides us the means to study vascular flows and perfusion inside bowel walls. The time-intensity curves (TIC) extracted from contrast clips are an efficient instrument of quantification, thus being also implemented in our research.

II. THE STATE OF THE ART

In [4] the authors analyzed the fluorescent images of colonic tissue based on textural parameters derived from the Grey Level Cooccurrence Matrix (GLCM), in order to distinguish the colonic healthy mucosa versus adenocarcinoma. A modified version of Multiple Discriminant Analysis was used for dimensionality reduction, such that only four final features resulted. The implemented Linear Discriminant Classifier provided 95% accuracy. In [5] the authors used the Grey Level Cooccurrence Matrix (GLCM), together with morphological features (shape, orientation), in order to characterize the malignant and benign tissues from biopsy slides of patients

Improving the textural model of the hepatocellular carcinoma through multiclass division using clustering methods

D. Mitrea[1], S. Nedevschi[1]
[1]Computer-Science Department
Technical University of Cluj-Napoca,
Cluj-Napoca, Romania
Delia.Mitrea@cs.utcluj.ro, Paulina.Mitrea@cs.utcluj.ro

M. Socaciu[2], M. Lupsor[2], R. Badea[2]
[2]Department of Ultrasonography,
3rd Medical Clinic,
Cluj-Napoca, Romania
socacium@yahoo.com, rbadea2003@yahoo.com

Abstract — The hepatocellular carcinoma (HCC) is the most frequent malignant liver tumor. The only reliable method for HCC diagnosis is the biopsy, but this is invasive, dangerous. We aim to develop computerized, non-invasive methods for HCC characterization, recognition, and evolution monitoring. We defined previously the imagistic textural model of HCC, consisting in the relevant textural features for HCC characterization and in the specific values associated to each relevant textural feature. We refine the imagistic textural model of HCC, by dividing this tumor in subclasses, based on its textural characteristics. We aim to find an improved characterization of HCC, concerning the tissue structure corresponding to each evolution phase. Several clustering methods are used in order to perform subclass division in the best manner. The relevant textural features will be determined for each subclass and also the effect of the multiclass classification techniques will be analyzed for the purpose of classification performance improvement.

Keywords – hepatocellular carcinoma (HCC); the imagistic textural model of HCC; HCC subclasses; clustering methods; performance assessment

I. INTRODUCTION

The hepatocellular carcinoma (HCC) is the most frequent malignant liver tumor (75% of liver cancer cases), besides hepatoblastoma (7%), cholangiocarcinoma and cystadeno-carcinoma (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 applied on ultrasound images. The texture is a very important visual feature, 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, of their specific values for the characterization of HCC, based only on information extracted from ultrasound images, and of the possibilities to obtain an optimal imagistic model of HCC 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. We previously gave the definition of the imagistic textural model of HCC [5], as consisting in the most relevant textural features able to separate the HCC tumor from the visually similar tissues (cirrhotic parenchyma, benign tumors) and in the specific values associated to the relevant features (mean, standard deviation, probability distribution). Concerning the selection of the relevant textural features, specific methods such as univariate density modeling through gaussian mixtures, correlation-based feature selection, Bayesian networks provided the best results. The Multilayer Perceptron and Support Vector Machines classifiers leaded to the highest recognition performance. Combinations of the feature selection methods, feature extraction methods, as well as metaclassifiers were also successfully experimented in order to improve the imagistic textural model [5]. The assessment of the evolution phase of the HCC tumor is also an issue of major importance. In this work, we elaborate an appropriate method in order to identify the HCC subtypes using the values of the textural parameters. Clustering methods, such as k-means clusters and X-means clustering are applied for this purpose. The most relevant textural features that correspond to the separation of the subclasses, as well as to the differentiation between HCC and the cirrhotic parenchyma on which HCC has evolved, are also determined.

II. THE STATE OF THE ART

The most frequently used methods in the field of texture-based characterization of the malignant tumors are the Grey Levels Cooccurrence 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

CARDIONET – A Distributed E-Health System for Patients with Cardio-Vascular Diseases

Gheorghe Sebestyen, Lorand Krucz
Computers department
Technical University of Cluj-N. TUCN
Cluj-Napoca, Romania
gheorghe.sebestyen@cs.utcluj.ro

Gavril Saplacan
Software Department
Applied Informatics Company , CIA SA
Cluj-Napoca, Romania
gsaplacan@yahoo.comd

***Abstract*— CARDIONET is a distributed e-Healh system developed on the latest IT&C technologies and standards, in order to improve the quality and responsiveness of medical care system for patients with cardio-vascular diseases. The system is meant to assure remote communication between patients and medical personnel and also service-based data exchange between different medical entities: hospitals, laboratories, healthcare assurance authorities, etc. The ontology-based approach facilitates seamless information exchange between different medical applications and allows reasoning and statistical analyses on the acquired medical data. The system includes mobile medical devices that assure on-line monitoring of patients' critical parameters.**

Keywords - e-Health, telemedicine, cardio-vascular diseases, patient monitoring

I. INTRODUCTION

Most national medical systems around the globe are struggling to increase the quality and responsiveness of medical services provided for their patients, under strict financial restrictions. In many diseases, including cardio-vascular ones, the time in which a patient is reaching and getting medical assistance may be crucial for the outcome of a medical episode. In today's medical practice too much time is spent in waiting rooms and even in hospitals. Most of these issues may be solved with the intensive use of existing information and communication technologies.

A distributed medical information system based on Internet technologies, which allows remote interaction between patients and medical personnel, can reduce the number of face-to-face visits and consequently the time spent by the patient in accessing medical assistance. Remote patient-medic interaction is recommended mainly in case of chronicle diseases (e.g. cardio-vascular ones) where a periodic and continuous adjustment of treatment is needed.

Through a better resource management such a system may also reduce the medical assistance costs and even the medical personnel's time. Patients may access medical services regardless of the physical distance between them and medical entities.

Furthermore, with the use of some low-cost mobile medical devices the doctor can follow and analyze on-line the critical medical parameters (e.g. EKG, pulse, blood pressure, temperature) of a patient being at home.

All these facts and observations motivated the development of the CARDIONET system, a distributed eHealth system dedicated for patients with cardio-vascular problems. The project, developed by a consortium of ITC and medical specialists, synthesize the latest developments, and practices in e-Health and telemedicine at national and international level. The main goal was to develop a flexible and scalable medical application model that may be adapted for different medical entities and pathologies and which can be the building block of a distributed medical system.

In the last decade similar initiatives and projects have been promoted in many European countries and in US. One of the most known projects is EPI-MEDICS [1], implemented through a collaboration of researchers from France-Italy-Sweden; in this project a miniature portable electrocardiograph was developed (Personal ECG Monitor), with 3 derivations for the detection of precocious miocardic coronary diseases and various arrhythmias

.

TOPCARE [2] is an example of a telemedicine platform that offer the tools needed for the remote surveillance of chronicle patients. The MOBI-DEV [3] project developed in Greece in a European partnership, tried to assure medical personnel's access to the databases that contained patients' information by means of wireless mobile technologies. The TELEMEDICARE project coordinated by Norway performed a complex 24/7 surveillance of patients equipped with biosensors. The SAPHIRE project aims to develop an intelligent healthcare platform integrating the wireless medical sensor data with hospital information systems.

Most of these projects are focused on some particular aspects of the cardiovascular patients' monitoring, such as: connectivity, responsiveness, reaction to some treatments, etc. without giving a global and patient-centric solution.

Therefore our research goal was to propose and implement a global medical system that integrates all the medical entities involved in a healthcare system (family doctors, hospitals, laboratories and social services) and facilitates seamless and transparent access to medical services through IT&C technologies. Taking into account the complexity of such an attempt, as a first step the patients with chronic cardio-vascular diseases were considered. But with

Integrated Informatic System used to Optimize the Diagnostic and Therapeutical Evaluation of the Patients with Chronic Myeloid Leukemia

Cristian Lelutiu[1], Sergiu Nedevschi[1], Luminita Lelutiu[2], Anca Bojan[2]

[1] Technical University of Cluj-Napoca, 28 Memorandumului Str., 400114 Cluj-Napoca, Romania
[2] Cancer Institute "Ion Chiricuta", 34-36 Gh. Bilascu Str., 400015 Cluj-Napoca, Romania
E-mail(s): Cristian.Lelutiu@cs.utcluj.ro

Abstract **— Chronic myelogenous leukemia (CML), is a myeloproliferative neoplasm defined by the presence of the BCR-ABL fusion gene produced by the translocation between chromosome 9 and 22 - t(9;22) (q34;q11). CML accounts for 15-20% of leukemias in adults, and as such is one of the most common leukemias. When untreated, its natural history is the progression to an acute leukemia (blast crisis) after a prolonged chronic phase. Although the recent development of inhibitors of BCR-ABL tyrosine kinase activity have dramatically altered the clinical course of CML,**
an important role has the management of the medical data, giving three main abilities :

1. **automatic extraction of suspicious data**
2. **early and accurate diagnostic and prognostic**
3. **good therapeutical evaluation and monitoring**

The current paper presents an integrated informatic system that permits the collaborative work of many specialists : general practitioners, clinicians, pathologists, laboratory specialists etc. The goal of this informatic system is to give the possibility of an early treatment for this category of patients and, through a good monitoring, to increase the remission period. This informatic system consists in four types of bar code driven modules, corresponding to the four health assistance system types : GP, laboratory/policlinic, hospital and high specialised clinical hospital. A very important feature is the interoperatibility with other health applications, using the HL7 (Health Level 7) message exchange protocol. The system is based on international standards (ICD10, LOINC), offering tools for remote interactions between all the entities involved in this process : pacients, doctors, automatic medical devices etc.

Keywords - Chronic myelogenous leukemia (CML) **• Chronic granulocytic leukaemia• *Philadelphia chromosome, BCR-ABL* • *CML prognostic models (Sokol score,Hasford score)* • *Use of cytogenetics* • *BCR-ABL point mutation* • *CML, Imatinib (Gleevec) resistance* • *HL7(Health Level 7)***

I. INTRODUCTION

One important concept introduced by the modern medicine science is the so-called evidence-based medicine (EBM). Basically, this concept consists in the integration of clinical expertise with external clinical evidence. These information flows are unified to develop the Electronic Health Record (EHR). The implementation of the HER needs a good definition of data elements, data types, unit and other attributes. That means to use a standardization. Some important organizations that are involved in the development of medical information standards are :

- ICD10 – International Classification of Desease, endorsed by the World Health Assembly in 1990, coming into use in WHO Member States beginning 1994
- LOINC – Laboratory Object Identifier and Numerical Code, offering a general standard code system for all the laboratory analises in close relationship with the next standard [2]
- HL7 – Health Level 7, a standard offered by a non profit organization, developed for the exchange of medical and administrative data between different medical entities [1]

Using these standards permits us to create a full interoperability between a large variety of medical applications, by formatting the data so that it can be received and computed at once.

II. BACKGROUND

Every disease can be diagnosed and treated by covering standardized stages :

1. Analysis of evident and/or suspicious clinical signs and simptoms
2. Laboratory tests :
   a. Common tests – shows that the health of an organism could be affected (can be transitory or not)
   b. Specialised tests - the diagnosis is restricted to a system or organ
   c. Specific tests - indicates the specific disease or can point out a specific abnormal status of the organism
3. Diagnosis stage, involving possible supplementary specific tests (imagistic, functional tests etc.)
4. Treatment / therapeutic stage
5. Monitorisation of the pacients, in case of :
   a. Diseases with possible recurrences
   b. Chronic diseases
   c. Possible side effects of the treatment
   d. Unexpected evolution of the pacients

There are four levels of health care services :

1. Family medicine, also called primary health care services system, performed by the general practiti-

Portal with Information and Evaluation Resources for Online Libraries

Alexandru Balog, Grigore Bădulescu, Revişoara Bădulescu, Ana-Maria Borozan
National Institute for Research and Development in Informatics, ICI
Bucharest, Romania
(alexb, grigoreb, revibad, marika)@ici.ro

Abstract - **The paper presents the portal with information and evaluation resources for online libraries developed by a consortium formed by the National Institute for Research and Development in Informatics, ICI Bucharest, the University of Bucharest and the Romanian Academy Library in the frame of a project from the National Research Plan Partnerships "Quality and performance evaluation of online libraries (LibEval)". LibEval portal is a unique access point to resources and information related to online libraries evaluation. The portal offers access to the existing best practices, applications, information and knowledge databases specific to online libraries for a large community of specialists and management personnel from libraries, researchers and experts in the domain of evaluation (especially library evaluation), academics and students, interested users. The user can access the portal content according to his access rights and the specific security level of the resource (document, link to a document or information) or event (news, scientific events, publications, courses specific to the domain etc.), can download a document and can collaborate to the content development with new resources and/or information using the existing collaborative tools.**

Keywords: online library, electronic library, digital library, library performance assessment, quality of electronic services for libraries evaluation, assessment methods.

I. INTRODUCTION

In the last decade important steps where made for the development of many digital libraries, of the digital content, of the technological and communication infrastructure and of the standards that will permit the operation of a digital library. Therefore, starting from the existent experience, the development of new methods and approaches for the evaluation of the quality and performances of a digital library is necessary.

The LibEval (Quality and performance evaluation of online libraries) project has as objective the research, design development and testing of a system with models, methods, software solutions and innovative services that will be used to assess the performance and quality of online libraries. The portal with information and evaluation resources for online libraries LibEval developed in the frame of the project offers access to the existing best practices, applications, information and knowledge bases specific to the domain of online libraries evaluation, to a large community of specialists like professionals and management staff of libraries, researchers and experts from assessment domain, especially from the libraries evaluation, academics and students, users interested by the field.

II. ONLINE LIBRARIES FEATURES

A sine qua non premise for a library evaluation, any type of library, is the identification of the creation and existence scope of the institution. The mission of a library is related to two key elements - **collections** and **users** -, organically linked by interoperational activities to administrate the transactions between them.

The management of the collections, users and relations between these categories respects the same criteria, regardless of the library type, traditional, non-automated, electronic, online, virtual and/or digital.

In the case of a library that produce, store and disseminates electronic content there is a third element that must by take into account, the **legislation** regarding the intellectual property rights.

Quality and performance evaluation of a library must stress the mode in which the institution responds to the specific requirements raised by each of the three elements.

The digital library is the most modern category from the online libraries and it is characterizing the institution which offers to users **automated bibliographical databases** and **digital content collections** and also **computer services**.

By the administration these three components in a coherent system, the users benefits from the retrieving services and information capitalization. The access to the database is obviously online. The system administration assumes content digitization operations but also information cataloging, storing, distributing, protecting and retrieving, typical operations for an online library.

The evaluation of the performances and quality of an online library looks after the success of the library by its impact on the independent user (in search for information) that navigates on the Internet and wishes to locate fast and easy the digital resources necessary for a specific activity/task. The evaluation must take into account the **services offered** by the library as support to the users tasks, **user satisfaction** and the **frequency of use** of the online library resources and services. The impact is given not only by the value of resources or services available to user but also by the extent in which the activities were determined impossible to be fulfilled without the use of online library.

The success of the library means that the user is satisfied by the resources developed and used by the library, by the value of the delivered information to create new knowledge, by the access alternatives and offered facilities, and that he will come back another time to use the online library services.

In the frame of the project there where inventoried and studied the existing approaches, practices and standards at international and national level regarding the libraries evaluation and especially of the digital libraries [1].

As noted by numerous specialists in the field (ex.: Bertot [6], [7], Saracevic [10], Kyrillidou [8], [9]), there are not standard definitions or approaches, strategies or practices regarding the evaluation steps. Each evaluation of a library

Digital Platform for Management of Information Dedicated to Research Projects

Doina Banciu, Dora Coardoş
Development in Informatics
National Institute for Research and– ICI, Bucharest, Romania
doina.banciu@ici.ro, coardos@ici.ro

Abstract - **The paper presents part of the achievements of scientific research obtained under a research project that started in 2009. The order is management of information regarding research projects and their results. More precisely, building a digital platform for information and documentation that contains information regarding research projects launched under national research programmers. The digital platform is developed as a Web portal, able to manage big volumes of vary information, as well as, allow on-line access to a great number of users. The digital platform is a complex Web solution, integrating advanced technologies for saving and up-dating information, in order to offer interactively information to the users.**

Keywords - digital platform; CERIF standard; information management; information and documentation; communication; Web portal.

I. INTRODUCTION

It is unanimously recognized that, in the 21st century, digital information and information networks are the main engines in economical growth and social development.

The research environment and the media that disseminates the research results are profoundly transforming in regard with the new technologies in information and communication that allow new opportunities and modification to the research activities. New opportunities and new models allow consolidation in dissemination of research results.

Access to global scientific and technologic information resources is a necessity and management of those resources has become an economic and political challenge, in order to ensure unrestricted access for every citizen.

In the field of scientific and technologic information and documentation systems, the global situation is complex, with important particularities in some developed countries. One can not talk about the existence of national range systems that would concentrate within their databases the results of Research & Development projects financed from the national budget or from international programmes. One can rather talk about spread information that stays either with the elaborating party, or within databases that belong to electronic publications, or handled by organisers of dissemination events.

At European level, creating digital content and building databases for a better dissemination of information are encouraged through all initiatives, in order to allow users on-line access to cultural resources.

Complying with worldwide tendencies and EU recommendations, in Romania as well, it is very important to create and implement information systems for Research & Development in various fields, accessible through Internet, integrated into an informational network dealing with Romanian Research & Development, integrated, in turn, to the European informational network for Research & Development.

This means, first of all, identifying, collecting and standardising vary types of data that relate to one's research activity, some of them complying with certain standards requested by the unified structure of the database, defining the technical media and terminology check tools for information retrieval, as well as training of staff that would use the system.

A significant example of a dedicated informatics system in the field approached is the CORDIS portal that contains information afferent to the results of Research & Development projects conducted under concluded European framework programmes.

In our country, the information related to results of budgeted R&D projects is stored in databases managed by authorities that conducted the programmes. This information is not yet accessible to the public, the Romanian scientific community.

Because of this we intent to build a *"Digital platform for management of information regarding research projects in science and technology",* that would take this information, store it in the system's database and make it available to those interested, through a Web portal.

II. DESCRIPTION OF THE PLATFORM

Systems used in information and documentation process are backed by well-structured, relational databases, have interfaces equipped with search engines for these databases, and usually, the access (even to summary information) cannot be provided without on-line registration.

Since the project was proposed to achieve by the National Authority for Scientific Research (ANCS) and that one of its main objectives is to develop a pilot digital platform for information-documentation on national research programs and projects and their results, the main

Patent Repository Management as Support for Innovation

Ágnes Sebestyén-Pál, Gheorghe Sebestyén-Pál, Gavril Săplăcan
Computer Science Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Agnes.Sebestyen.Pal@gmail.com

Marius Bulgaru, Horia Giurgiuman
TCM Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania

Marius.Bulgaru@tcm.utcluj.ro

Abstract— **The results and ultimately the success of innovation and research activities are highly dependent on a well-founded documentation process and inherently on the availability and the use of quality information sources. As argued in this paper, patents have a great potential as such valuable information source, with the condition that their availability and use is enabled and promoted by an efficient patent management system. While there are several official patent databases offering online access to patent documents, there is a considerable lack of patent managing services and an integrated repository system to incorporate these services. This paper presents the model for a patent repository and describes its collection of services that address specific patent processing tasks. As such, the repository addresses the issue of patent availability through patent harvesting, and assists its users in the process of information retrieval through services like patent processing, patent searching and report generation.**

Keywords- patent repository; document harvesting; information retrieval; semantic search; document classification; innovation

I. INTRODUCTION

The results and achievements of the activities involved in the field of research are known to be challenging to plan and control, making the process of innovation difficult to predict. However, there are contexts and conditions that stimulate and favor the development of innovative solutions, one of these being the availability of relevant information sources. Consequently, an important requirement in the field of research and innovation is to have the right means to support the documentation activities, involving searching for and classifying information.

Moreover, considering the great amount of information and the variety of information sources available (in the form of on-line books, journals, scientific articles, web pages etc.), there is an urgent need for tools that enable efficient and time-effective searches, resulting in meaningful documents. Simply put, this need refers not only to *finding* information, but rather to finding *relevant* information and *reducing* the obtained information amount to one manageable by a single person.

The need for such a solution is reflected by the amount of research done in the field of information processing and document retrieval. An overview of the relevant directions, strategies and fields of research is provided in [1], presenting approaches like vector-space models[2], natural language processing approaches, probabilistic models[3], latent semantic analysis[4] or neural networks.

An important issue that still remains open refers to the development of an integrated environment that combines the results of these different research areas into one composite solution and offers a platform that supports all representative processes, starting from information and document harvesting to document classification and result-synthesis through report generation.

The purpose and objectives of our work closely relate to [5] in that both approaches consider patents an efficient information source and promote the idea of using these documents for providing assistance in the innovation process. However, while Heusch focuses on a specific area (namely gearing systems) and developing a framework that automatically generates new ideas based on existing solutions, the current work addresses the more general problem of defining a set of services that provides a more efficient access to patents and a considerable support for document processing.

The rest of the paper is organized as follows: Section II discusses the potential of patents as relevant information sources and the issues related to ensuring patent access and availability. Section III introduces the proposed Pro-Inova Platform for Innovation, describing the purposes and functionality of the core set of services supported by the solution. This perspective is complemented by an architectural view that refers to the modules that implement the solution and details their requirements, responsibilities and interaction. The experimental results are discussed in Section IV and the paper concludes with the ideas presented in Section V.

II. PATENTS AS INFORMATION SOURCES FOR INNOVATION

One relevant decision in developing an integrated innovation platform refers to the selection of the appropriate

An Interactive System for the Grammatical Analysis of Written Texts in Romanian

Manuela Mihăescu
Sanda Cherata
Dina Vîlcu
Cornel Vîlcu
Babeş-Bolyai University
Cluj-Napoca, Romania

Augustin Terec
Software ITC Cluj Research Institute
Cluj-Napoca, Romania

Cornelia Melenti
The Technical University
Cluj-Napoca, Romania

Abstract — **The paper presents the realisation, within a research project, of an interactive system of grammatical analysis for texts written in Romanian. The two products realised as practical applications are presented here: a grammar checker for Romanian and an educational application with functions of assistance in teaching/ learning Romanian (as a foreign language).**

Keywords – computational linguistic; Romanian language; syntactic analysis; grammar checker; educational application

I. INTRODUCTION

The cultural and linguistic diversity, more and more pronounced due to encouragement and support of different countries in view of keeping their linguistic and cultural identity imposes, among other things, the refining of computer means of language processing. The extension of the European Union and the introduction of the Romanian language among the working languages imposed a thorough preoccupation for the development of more and more complex formal theories and descriptions, suitable for being integrated in complex computing applications.

The project presented here is related to this issue of language processing. Its purpose is to contribute to the formalisation of the Romanian language at a morphological, syntagmatic and syntactic level and also to the development and consolidation of the area of educational applications in the field.

II. THE SINTEGRO PROJECT

The SINTEGRO project (Interactive System of Grammatical Analysis for Written Romanian Language. Theoretical Model and Technology of Implementation - a research project of the type PNII, contract no. 11-022/2007) is realised in cooperation, within an interdisciplinary team, which includes Babeş-Bolyai University (UBB) from Cluj-Napoca as a coordinator, with the Research Institute SC Software ITC SA (SITC), Cluj-Napoca and the Technical University in Cluj-Napoca (UTCN) as partners. The site of the project is to be found at the address http://sintegro.rolingva.ro.

Included in the National Programme of Development P4, Partnership in Priority Areas, PNII, with the National Centre of Programme Management as contracting authority, the project falls in the research direction D1 (Technology of Information and Communication), with the specific objective *1.2.6 Computer systems for administration and development of interactive digital content (e-content).* (Manager of project: associate professor PhD Emma Morita – for the period 2007-2009, and teaching assistant PhD Manuela Mihăescu for the period 2009-2010).

The purpose of the team is to realise through this project an interactive system of grammatical analysis for written texts in Romanian. The practical applications that have been realised (and are close to being finalised) are: a *grammar checker* - SINTEGRO-VG, which can function as an independent application (in present, it is in the faze of experimental product) and an *educational application* - SINTEGRO-AE, equipped with a didactic interface and modules designed for teaching/ learning Romanian as a foreign language.

A. The research and development faze

The present research is based on a complex computer system for syntagmatic analysis of written texts in Romanian – SIASTRO - a system which was created and developed as part of an excellence project (ctr. no. 86 CEEX-II-03/2006, project manager associate professor PhD Emma Morita; coordinator: UBB, partners SITC and UTCN; Website: http://siastro.rolingva.ro). The results of this project materialised in the realisation of some original applications: LEXICON, which contains the vocabulary of Romanian language, with entries containing lexical, morphological and syntactic information; SIASTRO-AM - lexical-morphological and syntactic analyser and ETR – complex computer system for the extraction of terms from specialised texts written in Romanian.

The present project continues the research realised by the same team (UBB, SITC and UTCN) and introduces a new dimension in the linguistic analysis, namely the approach of the syntactic level, more exactly, analysis at trans-syntagmatic level.

B. Objectives

Concerning the general objectives, the main purposes of the project are 1) to enrich the theoretical resources that treat the Romanian language, by the formal description of a part of its syntax, and 2) to develop the area of computer applications for processing the Romanian language.

More exactly, the purposes of the project were:

- the formalised description of the structures involved in the grammatical checking of Romanian texts;
- the creation and the implementation of an experimental model of grammatical checker for the Romanian texts;

Scalable Ontology Processing in the ArhiNet System

Ioan Salomie, Razvan Lar
Department of Computer Science, Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Ioan.Salomie@cs.utcluj.ro

***Abstract*—This paper presents a scalable method for storing and processing an ontology with the use of a relational database. The ontology instance data is persisted in the database and the associated reasoning processes are reimplemented to leverage the relational structure. SWRL rules are translated into SQL stored procedures, instance realization is implemented by additional rules and SPARQL interrogations become simple SQL queries. To increase performance, the SWRL inference process is incremental, as only the inferences triggered by the latest ontology change are generated. Additionally, common SWRL rule fragments are identified and the associated data is shared between the rules in order to minimize duplicated work. Our work was tested on syntethic data as well as in the context of the ArhiNet system on knowledge extracted from documents addressing the medieval history of Transylvania and proved proved to be a scalable and efficient solution for ontology storage and reasoning.**

***Keywords*—SWRL, SPARQL, ontology, triple store, RDBMS, inference**

I. Introduction

Ontologies are a structured, expressive and compact way of representing knowledge about a certain domain such that this knowledge can easily be reasoned about. However, the most popular ontology storage format are simple XML files which must be loaded entirely into memory in order to perform any sort of querying or reasoning. Furthermore, many of the reasoning processes need to generate additional auxiliary data such that processing an ontology file of a few MB can require up to a GB of memory. This situation is especially problematic if ontologies are used as the primary information representation format in the context of a larger system, as it happens in the case of the ArhiNet system [1].

ArhiNet is an integrated historical documents processing system. The aim of the system is to extract knowledge from historical documents, to conduct automatic and manual reasoning on the extracted facts and finally to provide convenient access to this information to end users. The information that is extracted from the historical documents is structured in a domain ontology. Because the system handles large numbers of documents, the domain ontology also has a considerable size, which makes manipulating it difficult. Though there are several existing reasoners that provide many of the desired ontology manipulation functions, these are not able to handle large ontologies gracefully. This situation, aggravated by the large number of processed documents requires a more scalable ontology storage and inference mechanism.

This paper presents the RDBMS based ontology storage solution implemented in the ArhiNet system. The class and property instances of the ontology are persisted in a triple store and the SWRL inference and SPARQL query processes are reimplemented to leverage the relational database system and to provide good performance despite the size of the knowledge base. Each SWRL rule is implemented as a SQL stored procedure and each SPARQL query is executed by a SQL query.

The rest of the paper is organized as follows: Section II presents related work. Section III details the reasons for using a RDBMS for persistence. Section IV introduces the SWRL rule inference engine that can work with the database stored ontology. Section V presents the optimizations implemented for the rule-inference process. A method for executing a set of rules is presented in Section VI. In Section VII ontology realization is detailed. Section VIII describes how SPARQL queries are answered using the information in the database. In Section IX experimental results and implementation details are given. Finally, Section X contains our conclusions and future work proposals.

II. Related Work

Relational databases have been used for storing ontologies before. Namely, RDF representations of ontologies can be persisted in triple stores implemented in relational databases. The Jena Semantic Web Framework [2] can convert an ontology to a RDF graph and then persist it to either an in-memory or a RDBMS based triple store. However, the entire ontology is stored in the database despite the fact that some ontology elements would be more suitable for in-memory storage. Furthermore, Jena does not take advantage of the more complex query facilities of the database and instead performs joins between sets of triples on the client. This design decision allows in-memory and in-database triple stores to be treated equally, but hurts performance and scalability.

Optimizing the execution of a set of SWRL rules in a database is similar to multiple query optimization. In [3], a multi-query graph is constructed and heuristically partitioned such that intermediate result sharing between the queries is most efficient. In [4] an "AND-OR" DAG that incorporates all the sharing alternatives is traversed and an optimal strategy is extracted. The DAG however has a branching factor exponential in the number of joins in a query, which makes it impractical for use with SWRL rules that consist of many joins. In [5], a set of materialized views are available and each issued query is matched against the data in the views such that as much existing information is reused instead of recomputing it. The matching is similar to finding common fragments between a new SWRL rule and a set of already inserted SWRL rules. The process relies on a filter tree which is traversed from the root to the leaves, where the individual views are stored. Descending in the tree is equivalent to adding more filtering conditions on the initial data and the

A Distributed Approach for OWL Ontology Management and Query Answering

Mihaela Dinsoreanu, Angela Minculescu, Gabriel Sandor, Iulia Gorcea, Vlad Turian
Department of Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Mihaela.Dinsoreanu@cs.utcluj.ro

Abstract— **Ontology management activities represent a prevalent feature of most OWL editing environments, which ensure the modeling capabilities needed for obtaining comprehensive query answers. This paper proposes a distributed ontology management approach embedded in a semantic e-content processing system. It also presents a detailed comparison between how two different physical persistence methods for an OWL ontology – in regular operating system files or in a relational database – affects user query processing.**

Keywords - ontology, ontology reasoning, rules inference, ontology inconsistency, ontology graph view, knowledge engineering

I. INTRODUCTION

The necessity of digitizing archival content justifies the design of an integrated system that is able to manage various documents in a digital format, extract relevant knowledge from those documents and create a repository for that knowledge with the goal of enriching it. The ArhiNet system [10] is such an integrated system for the development and processing of distributed, semantically enhanced archival e-content. It tries to manage large amounts of archival data efficiently, by adding an additional, semantic layer over a set of digitally processed documents. The existence of this additional layer allows the development of a semantic knowledge base in the form of an ontology that offers advanced data management capabilities and efficient query-based information retrieval. One of the main goals of the ArhiNet system is to create a repository for the relevant knowledge extracted from the archival documents, aiming to enrich and make it available to the users. Within the ArhiNet system two such repositories were created: a relational database and a domain ontology, both of them supporting reasoning services.

Two of the features that highlight the importance of using an ontology as knowledge repository are the subject of this paper: the ability to perform various management tasks that change or improve the ontology and the capacity of fast retrieval of both asserted and inferred data using ontology-based query answering. Our solution is based on an in-memory representation of the domain ontology (the ontology model), whereas the relational database solution relies on the database engine to execute the queries or the rule inference and thus can scale better since secondary storage is used.

This paper presents a distributed approach to ontology knowledge engineering that has been implemented in the ArhiNet system, encapsulating the tasks that correspond to a human actor (namely the knowledge engineer) and the tasks that are performed in the background by the system.

The rest of the paper is organized as follows: Section II presents related work. Section III describes the ArhiNet system and an overview of how our approach to ontology management was applied in its context. Section IV presents our method for ontology management. Section V describes a method for query answering using both the ontology and the database as knowledge repositories. Section VI shows implementation details and experimental results. The paper ends with conclusions and future work proposals.

II. RELATED WORK

One of the most widely used OWL ontology editors in the semantic web industry is the open-source Protégé editor [1], developed at Stanford University. Among other functionalities, Protégé allows users to (*i*) manage OWL files, (*ii*) edit class expressions, (*iii*) use reasoners and (*iv*) design and add new components. However, the Protégé editor lacks some ontology debugging features that are implemented by Swoop [2], another OWL editor, which is a product whose development was stopped in early 2006. The debugging features from Swoop are very powerful and include: axiom pinpointing, root error pinpointing and ontology repair. Among other Swoop features we enumerate: change management, ontology versioning and Annotea. Both editors are desktop applications, whereas our approach provides a distributed environment for the various ontology-related operations.

The query answering systems are also difficult to find and customize. One of them is Gruff [9], a triple-store browser that can answer SPARQL [6] queries and that can visually represent the connections between the resources contained in a store under the form of a graph. Among the query answering systems we also mention jOWL [11], a JavaScript-based SPARQL endpoint. jOWL has many drawbacks, such as lack of support for some important SPARQL constructs and difficulty in modifying the queried ontology.

Our approach represents a complete and integrated system with both ontology management and query answering facilities.

III. THE ARHINET SYSTEM

ArhiNet is an integrated system for the development and processing of distributed, semantically enhanced archival e-content. In what follows, we will briefly present a high-level overview of the system and then we will focus on the features of interest in this paper, namely the ontology management and query answering modules.

A. The ArhiNet System's Architectural Overview

The conceptual architecture of the ArhiNet system is composed of three layers (see Figure 1): the Knowledge Acquisition layer, the Knowledge Processing layer and the

Ontology-Guided Natural Language-Based Knowledge Retrieval Method for Historical Documents

Ioan Salomie, Tea Morar, Romeo Zdrite
Department of Computer Science
Technical University of Cluj-Napoca
26-28 Baritiu str., Cluj-Napoca, Romania
Ioan.Salomie@cs.utcluj.ro

Abstract—**This paper presents a natural language interface for querying a digital knowledge base of historical documents. The complexity, vagueness and ambiguity of natural language often cause statements to be insufficiently clear to a software agent. To overcome the issue, we designed an ontology-guided natural language interface. We propose a method for semantically narrowing ontology suggestions and word meaning disambiguation. Dynamically-created SPARQL queries are used in our proposed approach to obtain increasingly focused suggestions with respect to the ontology content. Compared to other natural language interfaces designed with multi-ontology, domain-independent support, our domain-specific solution increases the accuracy of the retrieved results. The method was validated on knowledge acquired from a set of one-hundred historical documents.**

Keywords-natural language query*; *ontology*; *query suggestion*; *word disambiguation

I. INTRODUCTION

The World Wide Web has become home to an increasing number of domain specific repositories for content previously stored on volatile or perishable materials. The digitized content can therefore be accessed over a computer network such as the Internet by a substantially higher number of people. E-Content generation can be successfully applied to historical documents to enable widespread access to otherwise fragile, rare or unique documents which contain valuable information about important people, places and historical events. Most domain specific repositories have been adapted for query languages of different expressive power. These repositories support query interfaces, however they force the user to be proficient in a specific query language.

In this paper we propose a natural language-guided approach to querying a digital repository of archival documents. A guided natural language query written in regular, spoken language places technical and non-technical users on equal footing. Our approach removes the need for anthropologists, historians and archivists to learn an artificial query language while allowing users to compose precise queries. The proposed solution was tested on the ontology of the ArhiNet system [9]. The objective of the project is to develop an integrated system for managing semantically-enhanced archival content. The resulting system allows querying of the semantically-enhanced digital content by human users or software agents. The content is retrieved from the knowledge base built from processing the annotated documents. Based on archivist documents and information about specific Transylvanian economical, social and geographical structures the ArhiNet system builds a domain ontology according to which the raw documents are semantically annotated. These annotations facilitate document retrieval through ontology based querying. The query results represent the information requested by the user as well as the documents in which the information was located.

The rest of the paper is organized as follows. In section II we introduce related work. Section III presents an overview of the ArhiNet system conceptual architecture. Section IV details the Romanian query grammar while Section V focuses on query suggestions and disambiguation support. In Section VI we address query answering and in Section VII we provide experimental results and implementation details of our solution as part of the ArhiNet system. The paper ends with our conclusions and future work proposals.

II. RELATED WORK

This section reviews related work in the domain of natural language information retrieval. There are two areas of interest to the topic under discussion: the wider domain of natural language querying (henceforth referred to as NL querying) and the sub-domain of guided NL querying.

In the field of NL querying, PowerAqua [1] provides a solution for natural language interrogations with multiple-ontology support. A linguistic component analyses the NL query and associates it to a <term, relation, term> triplet discovered by a Triplet Similarity Service [1]. The identified elements will become part of a RDF triple to be used in querying all the ontologies discovered by PowerAqua's PowerMap [1] algorithm. PowerMap is defined as a hybrid knowledge-based matching algorithm which was utilized to translate NL terminology into ontology-compliant terminology [1]. Semantic filtering is used to determine which element will provide the answer type to the query.

Similarly to PowerAqua, AquaLog [2] receives as input a NL sentence which it translates into query triples. The query triples are matched by a relation similarity service to ontology compatible triples which are sent to the inference engine to retrieve an answer. AquaLog differs from PowerAqua in that it asks for disambiguation when a match between a query triple term and an ontology triple term cannot be found. If the system cannot differentiate between candidate concepts for the ontology compatible triples the user must provide feedback. The user feedback is then stored by a learning component in a synonym knowledge base.

Querix [3] provides a NL ontology-querying interface. As in the case of PowerAqua and AquaLog, Querix tries to match the words written in NL to ontology specific triples but

Semi-Automatic Method for Knowledge Acquisition from Historical Documents

Mihaela Dinsoreanu, Tamas-Albert Gasko, Adriana Dorina Ghirisan
Department of Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Mihaela.Dinsoreanu@cs.utcluj.ro

Abstract — **This paper addresses the problem of knowledge acquisition from historical documents. We present a method which semi-automatically extracts relevant information from Romanian texts written in natural language. The method adapts the Text2Onto framework to lexically and semantically annotate the text documents. The lexical annotations are obtained from a linguistic text analysis and processing. The semantic annotations result from applying a set of pattern/action rules on texts and are used to populate a domain ontology. The semantic annotations and the domain ontology can be used in knowledge retrieval. We tested and validated our method on a set of 100 documents addressing the history of Transylvania.**

Keywords - Lexical annotation; Semantic annotation; Modelling primitives; Probabilistic Ontology Model; Ontology.

I. INTRODUCTION

The famous author Pearl S. Buck once said: "If you want to understand today, you have to search yesterday". Yesterday refers to History, hence our great historical personalities and their actions, represent valuable and useful information for our own knowledge. The only source of such data is represented by the historical documents. To prevent the deterioration of the living and original proof of the past events, archival documents should be digitized. Moreover, in the current context, people are more familiar with sources in electronic format. Having digitized documents, Semantic Web based methods allow users to find relevant information easier. The ArhiNet system [1] is a solution for generating and processing semantically enhanced eContent, and allows users to find relevant information by means of ontology-guided natural language queries.

This paper presents a semi-automatic knowledge acquisition method which we integrated in the ArhiNet system. Starting with a core domain ontology and historical texts, this method lexically and semantically annotates the documents' content. The obtained annotations are used to enrich a historical domain ontology which is further used in knowledge inference and retrieval. To achieve these objectives, we adapted the Text2Onto [2] techniques to process Romanian text documents.

The rest of the paper is organized as follows. Section II presents related work, while Section III illustrates the conceptual architecture of the ArhiNet system and the associated workflows. Sections IV and V detail the method used for extracting the relevant information from texts and the method for enriching the domain ontology. Section VI presents a case study which illustrates how the knowledge acquisition method is applied in the context of Transylvanian historical documents. Section VII concludes the paper and presents future work proposals.

II. RELATED WORK

Some of the most important knowledge acquisition techniques presented in the research literature are OntoLT [3], OntoLearn [4] and Text2Onto.

OntoLT is available as a Protégé [5] plugin. OntoLT processes a collection of annotated documents in XML format that were previously obtained using WordNet [7] and the Stanford Parser [8]. The annotations enable the automatic extraction of classes and attributes which are used to build an ontology, according to a predefined set of mapping rules. If these rules do not totally correspond to the users' needs, the users can define their own rules, by means of a precondition language.

OntoLearn is a system that builds domain ontologies from documents. OntoLearn starts with the extraction of the appropriate domain terminology followed by creating sub-trees for the identified terms. For complex words, semantic interpretation is necessary, which is achieved by word meaning disambiguation, based on WordNet and a structural semantic interconnection (SSI) algorithm.

Text2Onto is a framework for ontology learning from text documents. For the linguistic processing of the documents it uses a General Architecture for Text Engineering (GATE) [6] pipeline that has as result lexically annotated texts. Based on the lexical annotations and on a set of Java Annotation Pattern Engine (JAPE) rules, semantic annotated documents are obtained. Finally a set of Text2Onto algorithms and combiners extract modeling primitives from the semantic annotations previously obtained, in order to enrich a core domain ontology. Unlike OntoLT and OntoLearn, Text2Onto computes a confidence degree for each identified object. Having the reliability of each fact expressed as a probability, helps the end users in making decisions regarding the facts' acceptance. Another advantage of Text2Onto is its ability to incrementally include new documents in the processed corpus without reprocessing the existing ones. These advantages are achieved by using a Probabilistic Ontology Model (POM), which is the representation in memory of the extracted information. Therefore POM stores all the probabilities assigned to objects and is updated only at corpus' changes. Another strong point of Text2Onto is that POM contains the learned data at a meta-level, allowing translation in different ontology representation languages, such as OWL [9] and RDFS [10].

Our approach uses and adapts the Text2Onto framework to build a historical ontology from archival documents written in Romanian language that are related to the history of Transylvania.

Enhancements on Adaptive E-learning Systems

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

Abstract- **Adaptive e-learning systems represent a new paradigm in modern learning approaches. A key factor in such systems is the correct identification of the user learning style, such as to provide the appropriate content presentation to each individual user. Moreover, a continuous re-evaluation and classification is essential in assessing the progress made during the learning process, and ensuring the evolution towards to a better style. This paper presents a solution for identifying the initial user typology, based on static features. Moreover, it describes the first steps in the design of the adaptive component of an e-learning system, which considers measuring the navigation elements, such as virtual notes, navigation path and concept maps. We propose the employment of a clustering method to determine the different groups of learning typologies, corresponding to the theoretical learning styles present in the literature. The evaluation results suggest that clustering provides a better correspondence between the individuals and the learning styles than a previous classification we have performed with Bayesian Networks. Moreover, the discrepancies observed in the results can be eliminated by careful design of the psychological test which measures the initial user static features.**

Keyword: Adaptive e-learning, learning style, concept maps, clustering.

I. INTRODUCTION

Following current trends triggered by the evolution of the technology, the education process has started to shift from the traditional face-to-face instruction to more modern approaches, such as online education. However, there exist a series of shortcomings which hinder the efficiency of such modern strategies, the most important being that the quality of the teaching effort is deficient. As a consequence, current trends in this area focus on the design of e-learning systems that contribute to the improvement the user's performance during the learning process. The goal is no longer the acquisition of knowledge alone, but how to do it in the most appropriate manner for each individual.

This paper presents a new possibility for identifying the user typology in an adaptive e-learning system previously designed by our team. We propose the employment of a clustering method to determine the different groups of learning typologies, corresponding to the theoretical learning styles present in literature. The rest of the paper is organized as follows: section II presents briefly the theory behind learning styles; section III presents a taxonomy of adaptive e-learning and different mechanisms we have employed in our e-learning system to achieve adaptation. Section IV presents the new proposed method for identifying the user type in the intelligent module and discusses the results of the evaluations. The concluding remarks and future directions are presented in section V.

II. LEARNING TAXONOMY

The main reason for using learning style theories in education is the improvement of the students' academic learning. This goal can be achieved by helping students comprehend the weakest and strongest points of their cognitive and meta-cognitive strategies. Data from learning styles inventories can be used by the students to observe and change their behavior, to monitor and choose the right learning strategies for the specific educational context. Following critical analyses of empirical studies, [1] we decided to use Vermunt's Learning Styles model [2]. This model is one of the most renowned because it includes: study motives, epistemological opinions, cognitive and meta-cognitive strategies used by the students in their learning process.

Through the Inventory of Learning Styles, Vermunt 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 of university students. It is experimentally 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.

There are a growing number of studies which propose the dichotomy between deep and surface learning. In the first case learners seek meaning through relating concepts and critical thinking about them and are intrinsically motivated. On the other hand, surface learners passively receive the information and try to memorize it, being extrinsically motivated. Regarding the influence of the learner's style on the elaboration of the concept map, Pearsall, Skipper and Mintzes [3] argue that deep learners make more complex concept maps. Also Carnot, Dunn and Cañas [4] find that deep learners locate information more quickly in a hypermedia course designed with concept map navigation support.

We have applied the Romanian version of the Vermunt Learning Style Inventory to 304 students from the Teacher Training Department in a Romanian Technical University. 118 students were female and 286 were male. The mean age of the students was 20 years. We employed the 100-item version of the Vermunt's Inventory of Learning Styles (*ILS*) [2]. The inventory consists of two parts, A and 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* has two parts: B1,