Integrated Web Services Platform for the facilitation of fraud detection in health care e-government services

A. Tagaris[1], G. Konnis[1], X. Benetou[1], T. Dimakopoulos[2], K. Kassis[2], N. Athanasiadis[3], Stefan Rüping[4], Henrik Grosskreutz[4], D. Koutsouris[1]

[1]Biomedical Engineering Laboratory, National Technical University of Athens, Athens, Greece, [2]AGILIS Informatics and Statistics SA, Athens, Greece, [3]Intrasoft International SA, Brussels, Belgium, [4]Fraunhofer IAIS, St. Augustin, Germany

***Abstract*—Public healthcare is a basic service provided by governments to citizens which is increasingly coming under pressure as the European population ages and the ratio of working to elderly persons falls. A way to make public spending on healthcare more efficient is to ensure that the money is spent on legitimate causes. This paper presents the work of the iWebCare project where a flexible, on-line, fraud detection, web services platform was designed and developed. It aims to help those in the Healthcare business, minimize the loss of funds to fraud. The Platform is able to detect erroneous or suspicious records in submitted health care data sets, ensuring homogeneity and consistency and promoting awareness and harmonization of fraud detection practices across health care systems in the EU. Critical objectives included, the development of an ontology of health care data associated with semantic rules, implementation and initial population of an ontology and rules repository, development of a fraud detection engine and implementation of a data mining module. The potential impact of this work can be substantial. More money on healthcare mean better healthcare. Living conditions and the trust of citizens in public healthcare will be improved.**

***Keywords* –fraud detection engine, ontology of health care data, rules repository, data mining.**

I. INTRODUCTION

According to NHS' Counter Fraud Service, EU healthcare expenditure is estimated to € 1 trillion per year [18]. EHFCN estimates that 3%-10% of these expenditures are lost to fraud [15]. Considering that EU members invest in healthcare each year between 3%-19% of their GDP and that by 2050 healthcare expenditures will reach between 6% to 7,6% of GDP from 5,3% in 2000, healthcare fraud will become costlier in the future. Hence, it is not surprising that combating fraud has gained interest in recent years and justifiably so, as it has demonstrated reduction in losses by as much as 45% in the case of NHS in the UK.

The iWebCare project analysed the prescription and purchasing processes of two public healthcare organizations (i.e. RBH and TSAY) and identified four major requirements. The counter fraud ICT tool must:

1. Codify organizational knowledge regarding fraud held by experts.
2. Automate the application of this counter fraud knowledge in order to harden existing business processes again fraud.
3. Better understand suspicious behaviour
4. Discover unknown types of fraud.

The Integrated Web Services Health Care Fraud Detection Platform is an advanced data analysis tool which combines data mining and rules based validation to detect suspicious (either erroneous or fraudulent) data submissions. It coordinates a collection of web services which are organized in three layers. The iWebCare platform follows a 3-layered architecture consisting of the web services (i.e. presentation), business and persistence layer. These layers implement the user interface, business logic and data storage respectively.

Web Services Layer

Consists of two types of services; those specific to the domain of Health Care Data Validation which includes: submission of data for inspection, submission of rules, submission of entity definitions, submission of erroneous data for training of the self learning module, request for rules, request for ontology information, etc; and supporting services such as user authentication, data encryption, document signing, etc.

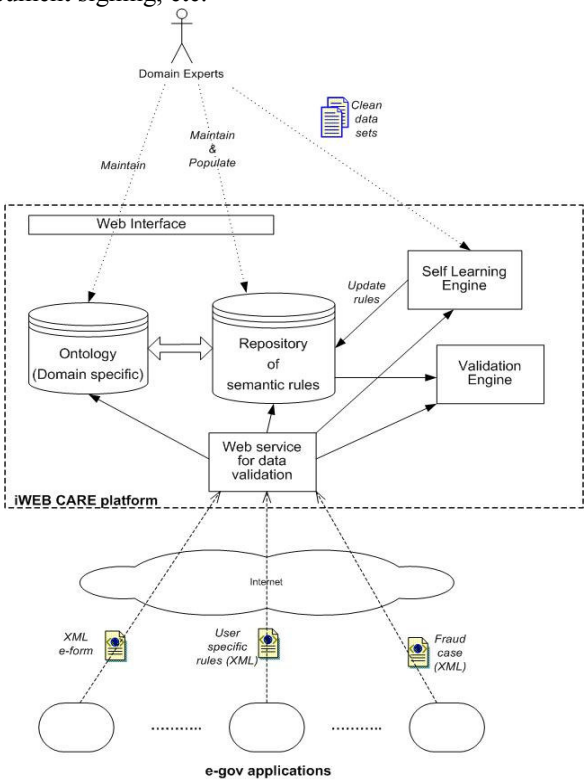

Fig. 1. iWebCare Overall Architecture

All web services were implemented using APIs provided by the Java 2 Platform, Enterprise Edition (J2EE). The use of this Java platform allows for operating system independency

as well as conformance to the Web Services and related standards such as SOAP [5] for the data messaging, WSDL [6] for the web service description (to be programmatically accessible from other applications), or UDDI [7] for registering the iWebCare services in web service registries. The specifications and guidelines of the Basic Profile of the Web Services Interoperability Organization (WS-I) were followed. Overall the web services were designed to interoperate as a Service Oriented Architecture (SOA) [8] system. The standard used for the actual orchestration of the services was the Web Services Business Process Execution Language (WSBPEL or BPEL) [9] from the OASIS standardization body.

Business Layer: This layer consists of two modules; the Validation Engine and the Self Learning module.

The Validation Engine is responsible for validating a submitted dataset message. A set of validation rules are applied to the dataset, according to the ontology information it contains. For that reason, the module must have access to the persistence layer to reach the required rules and ontology information.

The Self Learning module is responsible for identifying new semantic rules for entities and relationships between entities. Users enter into the system data with a known fraud status. The system utilizes data mining technologies to generate new rules which it stores for later use.

Persistence Layer: The main pieces of information that need to be persistent in the system are the:

a) *Ontology information:* Metadata of the healthcare domain entities, as well as the interrelations between them.

b) *Rules information:* Metadata related to suspicious values or combination of values for the entities of the ontology.

The persistence layer is accessed via standard JDBC [10] calls while all JDBC calls make standard SQL requests to the RDBMS [11]; thus RDBMS vendor independency is ensured.

II. DESIGN OF THE DOMAIN SPECIFIC ONTOLOGY

Ontology [3][4][13] defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts and relations among them. The main reasons for employing an ontology are to facilitate:

a) A common understanding of the structure of information among people and/or software agents
b) Reuse domain knowledge
c) Clarify domain assumptions
d) Separate domain knowledge from the operational knowledge
e) Analyse domain knowledge

The Artificial-Intelligence literature contains many ontology definitions [12][14]. Many of these are contradicting. For iWebCare's purposes, ontology is considered to be a formal explicit description of concepts in a domain of discourse (classes), properties of each concept describing various features and attributes of the concept (slots) and restrictions of slots (facets). Ontology together with a set of individual instances of classes constitutes a knowledge base. In reality, there is a fine line where the ontology ends and where the ontology base begins.

III. VALIDATION ENGINE

The validation engine [1] is mainly responsible for applying validation rules in the submitted structured Health Care (HC) datasets. The HC Datasets XML Schema is based in eGovML with minor changes to support Prescription details and more than one record in the same XML File. The main objective of the "eGOV initiative" was the provision of an open, extensible and scalable platform for realizing online one-stop government [23]. Each validation rule can be defined as the declaration of the domain, definition of a variable or of a combination of variables. To give some examples, Daily Drug Dose or the total cost of a medical examination is a positive integer number; the combination of drugs prescribed for a certain disease take certain values, etc. In other words, validation rules define domains of individual variables and place restrictions on the Cartesian products of individual variables' domains. Validation rules are declared to and accessed by the validation system as sets of variables and their domains and not as statements of relationships. In this way, validating a dataset simply means checking whether variables take allowed values or not.

The above mentioned functionalities are delivered through different modules. Each module is the object of a separate task, while the integrated engine was assembled and tested in a separate task, as described in the following paragraph:

Ontology and Rules retrieval module

This module is primarily responsible for identifying and creating the dataset to be validated. A user that is authenticated by the system to have the administrative privileges to perform validation uploads a dataset with structured HC data. The variables contained in the HC dataset are associated with the ones included in the domain specific ontology. Next, the declarations of rules including all necessary metadata that are applicable to this dataset are fetched from the corresponding rules repository and properly interpreted in a machine understandable way. Finally, the enhanced dataset that contains the actual data, the ontology's related concepts and the definitions of applicable rules are created by this module.

Fraud detection (validation) engine module

This module is responsible for the actual inspection of as well as any required action upon the enhanced dataset. The engine is responsible for the actual application of the rules (selected and edited by the user) to the enhanced dataset. Moreover, the engine upon discovery of suspect or erroneous records informs the user according to the rules' behaviour.

Reporting Module

This module handles the generation and presentation of report. When the actual validation process is complete the system calculates user selected validation metrics (e.g. percentage of failed records, number of records failing a particular validation rule) and produces the necessary post-validation reports in a user friendly format.

IV. Description of the Data mining Module

The data mining module delivers the data mining capability [16],[17] enabling the iWebCare platform to discover new rules. The data mining module identifies irregular patterns in the data. These are presented to the users who provide feedback allowing the system to improve its results.

The overall objective of the data mining engine was to integrate suitable algorithms for statistical data analysis and data mining tasks in the iWebCare platform in order to update, optimize and extend existing fraud detection rules. In order to address, these issues, the following tasks were executed:

⇨ Learn unknown rules from error free or fraudulent data in order to cover cases of fraud that are not detected by the current rule base.

⇨ Update existing rules by checking existing rules against new data. The correctness of these rules can be validated, and existing rules can be adapted to new data, e.g. by updating the value of the fraud score that is returned in a certain situation. In the case of unsatisfactory performance of the new rule set as a whole, a re-start of the learning process can be triggered. This addressed the dynamic nature of the application.

⇨ Find unknown, but useful, valid and comprehensive patterns in the data that may be used by experts to uncover new types of fraud that previously have not been identified.

⇨ Provide and supply the total portfolio of rules to the repository and to the web services

Basis of this work is a stable rules repository and a data 'feeder' sub-module of the iWebCare information system.

Workflow of the data mining module

The data mining module is implemented based on the data mining toolkit RapidMiner (formerly YALE) [19], which is wrapped inside a web-service interface and coupled with the Condor High Throughput Computing Environment[1] to guarantee high performance data analysis. The workflow for the data mining module is organised as follows (fig 2):

1. Data Input: Submitted legitimate or fraudulent data is sent to the module
2. Data Mining: The module analyses the data in different ways in order to test and find deviations from the existing set of rules
3. Plausibility checking: Statistically significant deviations are summarized in an intuitive representation and shown to domain experts, which can decide on their plausibility and practical significance. Practically not useful rules are rejected.
4. The data mining module 'forwards' the approved rules to the rules repository for updating the rules set.

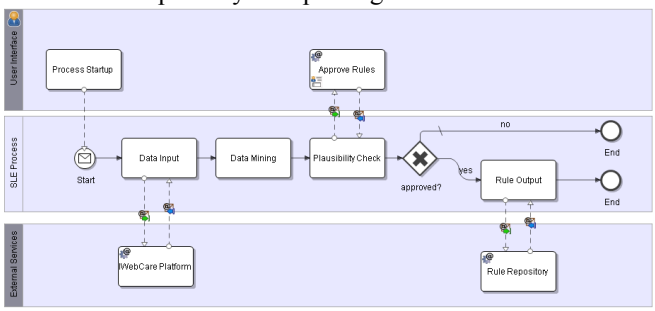

Fig 2. Workflow of the data mining module

In addition to the implementation of the data mining module, new data mining approaches have been developed to address the specific challenges of analyzing fraud data in the iWebCare approach

a) *Subgroup Discovery* [2] is a data mining approach that seeks to discover all statistically valid rules about an attribute of interest (such as the fraud) that can be found in a dataset. This is a contrast to standard approaches such as rule learning, which try to find only the statistically best rules to describe the single cases. However, in an approach such as in the iWebCare project, which is targeted at learning rules that are understandable and interesting to the domain expert, it is desirable to include all possible descriptions, such that the expert can select the most interesting or most practical ones [20]. The iWebCare data mining engine includes the currently fastest implementation of subgroup discovery [21].

b) *Similarity learning* seeks to automatically find a measure of how similar two entities are with respect to some aspect of interest (such as their fraud status). If a good measure of similarity has been found, it is easily possible to start from a single known fraud case and step by step investigate all similar cases for the same type of fraud. This local search for fraud effectively deals with the fact that in fraud detection it is often not possible to have enough cases of known fraud to extract a globally valid rule. In the iWebCare project, the potential of similarity learning has been demonstrated in the field of procurement fraud [22].

In summary, the data mining engine combines an easy extensibility with a wide range of data mining operations from the integration of a standard data mining toolkit with innovative algorithms that are targeted at important challenges of fraud detection which cannot be adequately solved with standard approaches, all wrapped into an efficient, easily portable distributed architecture.

V. Conclusion

The iWebCare platform was evaluated through two pilots and external experts who established a number of evaluation variables. These variables are divided in two broad

[1] http://www.cs.wisc.edu/condor/

dimensions: a) performance and quality (i.e. usability, reliability, efficiency, functionality, satisfaction) and b) business dimensions. The first category focused on technical characteristics of the platform. The latter considered the benefits to the client organization of using the iWebCare platform to detect potential fraud and looked at how much potential fraud the platform was able to identify, the worth of the potential fraud and the impact of the platform financially, technologically and organizationally.

The two pilots evaluated for the same variables, even though, they executed completely different scenarios. The first scenario focused on conflict of interest fraud in non-medical procurement processes within RBH/NHS (major cardiothoracic hospital). The second scenario examined fraud in the prescription processes of TSAY (Social Security Organization). The pilots used similar collection and evaluation methods in order to allow for comparisons between the pilots. The external experts were able to validate the testing process since the number of test subjects was limited as the platform audience is small.

The pilot was successful overall because it achieved its key objectives to conduct trials in order to validate the iWebCare service from the viewpoint of end-users. Users found the platform easy to use and the platform identified cases of potential fraud. The results of the self learning module were also welcomed.

The users highlighted their belief that the use of the platform on a regular basis would reduce administrative costs to support current processes and improved effectiveness by automating tasks to deal with large volumes of data which would be difficult to process without the use of the platform. The external experts collaborated user findings.

Although potential fraud has been identified through use of the iWebCare platform, further investigation is required from the fraud experts before fraud can be confirmed. Experts identified areas of improvement for the platform:

- ⇨ Extending the interoperability properties of the platform in order to link to external data sources
- ⇨ Rule editor

The overall evaluation and assessment of the iWebCare integrated web services platform have shown us that the platform is a very useful platform for detecting and fighting fraud not only for the healthcare domain but for any government and business domain.

In the future the applicability of the iWebCare platform is planned to be adopted by other domains related to public authorities and e-government such as e-Procurement, Registrar's offices, Public Transportation, Revenue Authorities, Customs and Insurances. The overall vision is to provide a powerful tool (iWebCare platform) to aid in the detection of fraud in domains and areas of vital importance that have a direct relation to the overall cost of business.

VI. ACKNOWLEDGMENTS

The work presented in this paper was funded by the European Commission for the iWebCare project under Grant FP6-2004-IST-4-028055.

REFERENCES

[1] Petrakos, Conyersano, Farmakis, Mola, Siciliano and Stavropoulos, "New ways to specify data edits", Journal of Royal Statistical Society, Ser. A, 167, Part 2, 294-274, 2004

[2] Willi Klösgen. *Handbook of Data Mining and Knowledge Discovery*, chapter Subgroup discovery. Chapter 16.3. Oxford University Press, New York, 2002.

[3] Oscar Corcho1, Mariano Fernández-López, Asunción Gómez-Pérez, Angel López-Cima, "Building legal ontologies with METHONTOLOGY and WebODE"

[4] Gómez-Pérez A (editor) (2002) "A survey on ontology tools", OntoWeb deliverable

[5] http://www.w3schools.com/soap/default.asp

[6] http://www.w3.org/TR/wsdl

[7] http://www.uddi.org/

[8] http://www.soa.com/

[9] http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsbpel

[10] http://java.sun.com/javase/technologies/database.jsp

[11] http://www.webopedia.com/TERM/R/RDBMS.html

[12] http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

[13] http://www.w3.org/TR/owl-features/

[14] Natalya F. Noy, Deborah L. McGuinness, "Ontology development 101: A guide to creating your first ontology", Stanford University, Stanford, CA, 94305

[15] http://www.ehfcn.org/declaration.html.

[16] Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning*, Springer, 2001

[17] Hand, Mannila, and Smyth, *Principles of Data Mining,* MIT Press, 2001

[18] http://www.nhsbsa.nhs.uk/CounterFraud.aspx

[19] Mierswa, Ingo and Wurst, Michael and Klinkenberg, Ralf and Scholz, Martin and Euler, Timm, "YALE: Rapid Prototyping for Complex Data Mining Tasks", Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006.

[20] Grosskreutz, Henrik, Rüping, Stefan and Wrobel, Stefan. "Tight Optimistic Estimates for Fast Subgroup Discovery". In: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Springer LNAI, 2008.

[21] Rüping, Stefan: "Ranking Interesting Subgroups". In: Proceedings of the 26th International Conference on Machine Learning (ICML 2009), Bottou, Leon and Littman, Michael (Eds.), Montreal, Omnipress, 913-920, 2009.

[22] Rüping, Stefan, Punko, Natalja, Günter, Björn and Grosskreutz, Henrik. "Procurement Fraud Discovery using Similarity Measure Learning". In: Transactions on Case-based Reasoning, 1(1), 37-46, 2008.

[23] Governmental Markup Language (GovML) for Online One-Stop E-Government http://xml.coverpages.org/ni2002-02-27-a.html