# Dynamic Terminology Enhancement for Integrated ECG Resources

Alexandra Kokkinaki, Ioanna Chouvarda, Nicos Maglaveras

Aristotle University of Thessaloniki, Thessaloniki, Greece

## Abstract

In this paper, we present Dynamic Terminology Enhancement Method (DTEM) to support enrichment and extensibility in a biosignal integration system called ROISES (Research Oriented Integration System for ECG signals), which integrates diversely encoded ECG signals and the corresponding annotation and metadata. The diverse datasources are homogenized through the mapping of their schemas to an ECG specialized global ontology (GO). DTE method combines UMLS rich terminology and machine learning techniques to first determine the suitability of a term to constitute global *ontology's class and secondly locate its position in GO's* hierarchy.

## 1. Introduction

Public Biosignal Resources, containing signals and annotations, can serve an important role in biomedical research and education, provided that an appropriate framework allows for population, querying and data/metadata access of these resources. In this scope ROISES framework was developed to enable the unified access to bio-signal databases and the accompanying metadata. It allows decoupling information retrieval from actual underlying datasource structure and enables transparent content based searching from multiple data resources with context filtering. ROISES provides a reconciled view of different ECG repositories through the use of ontologies, ECG domain standards and UMLS [1]. The system was designed to serve the complex requirements of the medical researcher, who studies the evolution of ECG parameter values, constrained by specific criteria. ROISES framework is also intended for the medical student who seeks exemplary annotated ECGs in order to self practice in the identification of the structural and pathophysiologic abnormalities recognized in ECG patterns.

ROISES allows for the unified access to the information encompassed in ECGs encoded in diverse medical standards, such as SCP-ECG [2] and EDF [3], and located in different data structures, such as databases and ontologies, by segregating the content and context of the signal. Biosignal context refers to information about patient demographics, diagnosis, recording equipment, researcher/investigator, etc., while annotation and interpretation information constitute part of the bio-signal content. The diverse biosignals are semantically unified in terms of GO, which was structured according to ecgML [4], a markup language for ECG representation. Additionally, GO's taxonomy encapsulates UMLS terminology pertaining medication, diseases, diagnosis and coding schemes, like Lead coding system originating from biosignal medical standards such as SCP-ECG .

Architecturally, ROISES consists a Local As View (LAV) semantic data integration system, which requires the global schema to be specified independently from the sources. The relationships between the global schema and the sources are established by defining every source as a view over the global schema [5]. ROISES includes its own mapping procedure, during which not all terms originating from the sources are mapped to terms in the global ontology. These terms, although marked as "unmapped" by the mapping procedure and are excluded from the query process, could possibly constitute significant query criteria and research parameters for the medical researcher. In this scope, Dynamic Terminology Enhancement Method (DTEM) is proposed as a semi automatic method to enrich and extend the global ontology by processing the previously defined unmapped terms. DTEM employs Data Mining techniques, particularly Machine Learning techniques, to initially determine the suitability of a term to constitute a class in the global ontology and secondly to define its position in GO's taxonomy.

## 2. Materials and methods

DTE method, which is a specialized ontology enrichment method, employs UMLS Metathesaurus to compare an unmapped term to corresponding UMLS concepts, while UMLS Semantic Navigator elicits the term's contextual information, i.e. its organization in the source vocabularies. Furthermore, DTEM employs machine learning techniques to initially determine the suitability of a candidate term to constitute an ontology class and secondly to locate its position in the ontology hierarchy.

DTE method accepts as input a series of terms, not mapped to any of GO's concepts during the mapping process. Following, the method determines the suitability of the candidate terms and their position in GO's taxonomy. The previously defined subtasks are performed

ISSN 0276–6574

as two separate but interdependent mechanisms, namely Concept Suitability Investigation Mechanism and Hierarchy Suitability Investigation Mechanism.

## 2.1. Concept suitability investigation mechanism

During Concept Suitability Investigation Mechanism, the candidate terms are entered in Metathesaurus search engine, to browse information concerning synonyms, their preferred name their Concept Unique Identifier (CUI). A simple search for a medical concept induces the following cases:

a) No results are returned
b) Only one term is obtained
c) More than one terms are acquired

Case a) denotes that this term is not related to any medical concepts in UMLS lexical database, so this term is rejected. Cases b) and c) lead to the extraction of additional features for every returned UMLS CUI. This set of features formulates vector $V_{F_i}=(f_1,f_2,f_j,\ldots\ldots f_N)$, where $f_j$ is the value of the jth feature, and N is the number of collected features, with N>0. Apparently, the problem of characterizing a medical term as suitable to form a class (1) or not (-1) is reduced to a binary classification problem. In this scope, Support Vector Machines were employed to classify the candidate terms as suitable or inapt to form ontology classes.

For this particular problem the Linear Function and the Radial Basis Function (RBF) were applied. The Linear Function is the simplest one, while RBF kernels, described by equation 1, give good generalization performance, need less hyper parameters than the polynomial kernel, and are considered as a generalization of the linear kernel function.

K(xi, xj) = exp($-\gamma$‖xi-*xj*‖^2) *for* $\gamma>0$ (equation 1)

The classification process requires the segregation of the datasets into training and testing datasets. The training dataset D, where D = ($V_{FiM}$, $y_M$), and M is the number of training data instances, f∈R and y∈ {-1,1}, which determines if a concept will form a class in GO's taxonomy (1) or not (-1). Given a testing dataset, T=($V_{F_k}$), k=1,2…L, where L is the number of data, an SVM classifier predicts the $y_i$ values. The elements in the feature vectors, which are numerical indicators of the semantic and syntactic distance between two medical terms, are defined as follows:

$f_1 \in \mathbb{R}^+$, is the Edit Distance between a Medical term and its UMLS synonym,

$f_2 \in \{0, 1, 2\}$where $f_2$ equals 0 when the Semantic Type/Group of the medical term and the UMLS synonym are irrelevant, 1 when they belong to the same Semantic Group, 2 when they share the same semantic type, and

$f_3 \in \mathbb{R}^+$, is the number of UMLS results returned for a medical term.

To evaluate the accuracy of a classifier, three measures are typically employed, namely the sensitivity (sens.), the specificity (spec.) and the accuracy, described by equations 2, 3, 4 respectively:

sens. = TP/(TP+FN) (equation 2)
spec. = TN/(TN+FP)- (equation 3)
accuracy. = (TP+TN)/(TP+FP+TN+FN)- (equation 4)

where, TP is the number of true positives, i.e. the positive examples correctly classified as positives, TN is the number of true negatives, FP is the number of false positives, i.e. the negative examples incorrectly classified as positives, and FN are the false negatives.

## 2.2. Hierarchy suitability investigation mechanism

HSI mechanism aims at incorporating the candidate classes along with their ancestor hierarchy, into GO's taxonomy. To accurately determine the appropriate insertion points, UMLS Knowledge Source Server (UMLS KSS) is employed to retrieve the list of indented hierarchies for the concepts' ancestors. A single medical term most probably yields more than one ancestor hierarchies originating from different source vocabularies.

Each tree hierarchy is expressed as an undirected graph as depicted in Table 4. For example, the medical term Hyperlipoproteinemias induced the hierarchies presented in the third column of Table 4, originating from the source vocabularies SNOMED CT and MeSH as depicted in the second column.

When querying UMLS KSS, it is important to limit results to those having a Source Abbreviation (SAB) equal to "MSH", "NCI" "SNOMEDCT" which specify MeSH (Medical Subject Headings), National Cancer Institute Thesaurus (NCI Thesaurus) and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) respectively, since their structure resembles GO's hierarchy more accurately. Furthermore, a new parameter was specified, namely Required Starting Concept, which represents the concept that will subsume the required hierarchy.

In this sense, HSI constructs K feature vectors for each medical term, where K is the number of UMLS results for each term. The feature vector of the $i^{th}$ term is $V_{Pi}=(p_1,p_2,p_j\ldots. p_S)$, and defines the attributes and characteristics of the resulted ancestor hierarchies. $p_j$ is the value of the $j_{th}$ attribute while S is the number of characteristics in the feature vector with S>0. The elements in the feature vectors constitute numerical indicators of the similarity among the UMLS returned hierarchies and a particular part of GO's hierarchy, as described below:

$p_1 \in \mathbb{R}^+$, is the number of UMLS returned results per dictionary,

$p_2 \in \mathbb{R}^+$, is the number of retrieved UMLS hierarchical nodes,

$p_3 \in \mathbb{R}^+$, is the number of common nodes among UMLS resulted hierarchy and GO,

$p_4 \in \mathbb{R}^+$, is the depth of the first occurrence of the Required Starting Concept,

$p_5 \in \mathbb{R}^+$, is the last common node in terms of depth in UMLS result hierarchy,

$p_6 \in \mathbb{R}^+$, is the last common node in terms of depth in GO hierarchy

$p_7 \in \{0, 1, 2\}$, where 0, 1, 2 correspond to results originating from “MSH”, “NCI” and “SNOMEDCT” respectively.

As previously demonstrated, a medical term may yield a number of UMLS ancestor hierarchies, which are organized in the respective number of feature vectors. $V_{P7}=(p_1,p_2,p_3,p_4,p_5,p_6,p_7)$. HSI mechanism aims at suggesting to the domain expert, the most suitable hierarchies to be incorporated in the global ontology. Distinguishing among suitable and not suitable hierarchies reduces the problem to binary classification. The classification was accomplished with supervised machine learning techniques, namely classification trees and Naïve Bayes classifier.

## 3. Results

DTEM method was evaluated after registering three biosignal sources to ROISES framework. The sources comprised a relational database for the management of patients with chronic heart disease, a collection of SCP-ECG files and a collection of EDF plus files. The application of the mapping procedure to homogenize the differently formatted sources with the Global Ontology (GO) yielded a number of unmapped terms which were entered as input to CSI mechanism.

### 3.1. CSI results

CSI mechanism establishes a connection to UMLS Knowledge Server, to extract the necessary features for each unmapped term, in order to construct Vector $V_{F3}=(f_1,f_2,f_3)$. To evaluate the mechanism, a specific example is demonstrated, the terms of which were extracted from the free text field Medical History of SCP-ECG encoded bio-signals and are depicted below:

[Colon cancer, insulin_coma, Hyperlipoproteinemia, ACVB (1x) ,Arterial hypertension, Rheumatoid arthritis, Nephrotic syndrome, Skoliosis, , Hypercholesterinemia, Megaloblastic anemia, Struma (euthyroid) , Obesity, Renal insufficiency, Type IIa , peripheral atherosclerosis, Vit. B12 deficiency, Diabetes mellitus, Gastritis, Hyperlipoproteinemia Type IV, Rheumatoid arthritis, acute_respiratory_alkalosis, Adiposis hepatis, Hyperlipoproteinemia Typ IIa, Hyperuricemia].

CSI mechanism organizes the UMLS extracted features per term, in the following order: Required Term → Semantic Type → UMLS term → Semantic Type → Semantic Group. The mechanism constructs none, one or more structures for each required term which are summarized in Table 1.

Table 1. Application of CSI mechanism in part of the group of requested terms.

| Requested Term | Result |
|---|---|
| Colon cancer | Colon cancer --> Disease or Disorder---> Colon Carcinoma ---> Neoplastic Process; ---> Disorders<br>Colon cancer --> Disease or Disorder ---> Malignant tumor of colon ---> Neoplastic Process; ---> Disorders |
| Arterial hypertension | Arterial hypertension --> Disease or Disorder ---> Hypertensive disease ---> Disease or Syndrome; ---> Disorders |
| Obesity | Obesity --> Disease or Disorder ---> Obesity Adverse Event ---> Finding; ---> Disorders<br>Obesity --> Disease or Disorder ---> Obesity ---> Disease or Syndrome; ---> Disorders |
| Renal insufficiency | Renal insufficiency --> Disease or Disorder ---> Renal Insufficiency ---> Disease or Syndrome; ---> Disorders |

To classify the suitable from the inapt classes, we applied Support Vector Machines (SVMs) with linear and Radial Basis Function (RBF) kernels. In this scope the terms are distinguished in testing and training datasets. The testing set contains 85 medical terms retrieved from UMLS Metathesaurus 76 being positive (suitable classes) and 9 being negative (inapt classes) with each medical term being represented by 3 features (see section 2.1). The training set consists of 28 medical terms 9 of which are labeled as negative and 19 as positive ones. Concerning the RBF kernel, grid search was applied to the training dataset, to select the best-fitting C and $\gamma$ parameters within a range of [-5, 5] and [-4, 0] respectively, with step 1 in the log scale. Parameters C and $\gamma$ were set to 2 and 0.125 respectively during the training procedure. To fairly compare the different kernels, the same training set was also employed to train the linear classifier. The generated models were then adopted to predict the target value of the data instances in the testing set. For these models, sensitivity, specificity and accuracy were employed to evaluate the method’s accuracy. As depicted in Table 2, the RBF classifier succeeds much higher accuracy, in terms of both sensitivity and specificity, than the linear SVM classifier applied to the same training and testing datasets.

Table 2. Sensitivity, specificity and accuracy of Linear and RBF kernels.

| Kernel | Sens. | Spec. | Accuracy |
|---|---|---|---|
| Linear | 0,97368 | 0,666667 | 94.1176% |
| RBF | 0,98684 | 0,777778 | 96.4706% |

## 3.2. HSI results

To classify the suitable from the inapt ancestor hierarchies, HSI mechanism employs machine learning techniques. The terms used to construct the classification datasets originate from data pertaining Medication, Medication Categories and free text Diagnosis and were derived from the previous step. The training dataset consist of 55 instances 49 being negative (-1) and 6 being positive (1) while the testing dataset comprise 61 instances, 51 being positive and 10 being negative. The machine learning techniques engaged are the classification trees (C4.5 and ADT (Alternating Decision Trees) algorithms) and Naïve Bayes. Table 3 summarizes the results concerning sensitivity, specificity and accuracy.

Table 3. Sensitivity, specificity and accuracy of Linear and RBF kernels.

| Algorith | Sens. | Spec. | Accuracy |
|---|---|---|---|
| ADTree | 0,8 | 0,727 | **91.803%** |
| C4.5 | 1,0 | 0,833 | **96.721%** |
| Naïve | 1,0 | 0,833 | **96.72** |

Applying C4.5 algorithm to HSI mechanism yields the results presented in Table 4. These taxonomies need the domain expert's validation before being incorporated in GO's hierarchy. The first column of the table depicts the candidate class and the CUI code extracted by UMLS. The second column shows the name of the UMLS source vocabulary (SAB), while in the third column the ancestor hierarchy of the candidate class is presented. This hierarchy will be subsumed under the corresponding Required Starting Concept in GO's hierarchy, i.e. class Disease after the domain expert's validation.

Table 4. Partial results from applying C4.5 algorithm to HSI mechanism.

| CUI/ CName | SAB | Ancestry |
|---|---|---|
| C0020476/ Hyperlipoproteinemias | SNOMED | Disease--->Metabolic disease--->Disorder of lipoprotein AND/OR lipid metabolism--->Disorder of lipoprotein storage and metabolism |
| C0020476/ Hyperlipoproteinemias | MeSH | Diseases (MeSH Category)--->Nutritional and Metabolic Diseases-- >Metabolic Diseases--->Lipid Metabolism Disorders--->Dyslipidemias--->Hyperlipidemias |
| C0740394/ Hyperuricemia | SNOMED | Disease--->Metabolic disease--->Inborn error of metabolism--->Disorder of purine and pyrimidine metabolism--->Disorder of purine metabolism--->Increased uric acid level |
| C0740394/ Hyperuricemia | SNOMED | Disease--->Congenital disease--->Inborn error of metabolism--->Disorder of purine and pyrimidine metabolism--->Disorder of purine metabolism--->Increased uric acid level |
| C0740394/ Hyperuricemia | SNOMED | Disease--->Hereditary disease--->Inborn error of metabolism--->Disorder of purine and pyrimidine metabolism--->Disorder of purine metabolism--->Increased uric acid level |
| C0028754/ Obesity | MeSH | Diseases (MeSH Category)--->Nutritional and Metabolic Diseases--->Nutrition Disorders--->Overnutrition |

## 4. Discussion

ROISES framework currently focuses on ECG, and it currently aggregates two bio-signal data formats (SCP, EDF) and two data structures (relational database, ontology). Extending the framework means adding more data to the already registered structures or/and encapsulating more biosignal standards to the framework. DTEM method supports the extensibility of the system in both ways. From a technical point of view the method fosters ROISES framework by not only extending the sources but also continually enriching its terminology. From a research point of view, updating and encompassing new data sources in the system, adds to the variability of query criteria, thus providing the final user with a dynamic and renewable system, which reinforces integrated research.

## References

[1] Unified Medical Language System®. Unified Medical Language System (UMLS). 01.11.2006 Available from: http://www.nlm.nih.gov/research/umls/

[2] OpenECG Consortium, "SCP-ECG standard last work document", 2005. [Online] Available : http://www.openecg.net/member/EN1064_lastworkdocument.pdf

[3] Kemp B**,** Olivan J. European data format 'plus' (EDF+), an EDF alike standard format for the exchange of physiological data. Clin Neurophysiol. 2003:114:1755-61.

[4] Wang H, Azuaje F, Jung B, Black N. A markup language for electrocardiogram data acquisition and analysis (ecgML). BMC Med Inform Decis Mak. 2003:3

[5] Lenzerini M. Data Integration: A Theoretical Perspective. 21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS02). 2002: 233-246.