COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 108 (2012) 453–466

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Searching biosignal databases by content and context: Research Oriented Integration System for ECG Signals (ROISES)

*Alexandra Kokkinaki, Ioanna Chouvarda, Nicos Maglaveras**

Lab. of Medical Informatics, The Medical School, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 1 February 2010
Received in revised form
28 December 2010
Accepted 12 February 2011

Keywords:
Biosignal database
Ontology
Semantic and syntactic mapping

ABSTRACT

Technological advances in textile, biosensor and electrocardiography domain induced the wide spread use of bio-signal acquisition devices leading to the generation of massive bio-signal datasets. Among the most popular bio-signals, electrocardiogram (ECG) possesses the longest tradition in bio-signal monitoring and recording, being a strong and relatively robust signal. As research resources are fostered, research community promotes the need to extract new knowledge from bio-signals towards the adoption of new medical procedures. However, integrated access, query and management of ECGs are impeded by the diversity and heterogeneity of bio-signal storage data formats. In this scope, the proposed work introduces a new methodology for the unified access to bio-signal databases and the accompanying metadata. It allows decoupling information retrieval from actual underlying datasource structures and enables transparent content and context based searching from multiple data resources. Our approach is based on the definition of an interactive global ontology which manipulates the similarities and the differences of the underlying sources to either establish similarity mappings or enrich its terminological structure. We also introduce ROISES (Research Oriented Integration System for ECG Signals), for the definition of complex content based queries against the diverse bio-signal data sources.

1. Introduction

Information and communication technologies have grown rapidly during the last decades and are vastly used in every aspect of human activity. In healthcare environment like primary care and hospitals, the most visible technological developments of the last thirty years are electronic health records, tele-medicine, bio-signal processing, medical image interpretation, and knowledge processing. The wide spread use of bio-signal acquisition devices in daily life applications, generated and continues to generate large amounts of bio-signals. Following the pervasive healthcare directions, access to bio-signals has been diversified and is no longer restricted to the hospital domain. Therefore, research resources are fostered and research community promotes the need to extract new knowledge from bio-signals towards the adoption of new medical procedures.

In the cardiology domain, the recording of the resting 12-lead ECG continues to be the most commonly used laboratory procedure for the diagnosis of cardiac disease. As a record of the electrical activity of the heart, it is a unique technology that provides information not readily obtained by other methods [1]. However, ECG bio-signals are not efficiently managed,

* *Corresponding author at*: Lab. of Medical Informatics, Faculty of Medicine, The Medical School, Aristotle University of Thessaloniki, P.O. Box 323, 54124 Thessaloniki, Greece. Tel.: +30 2310999281; fax: +30 2310999263.
E-mail addresses: alko@med.auth.gr (A. Kokkinaki), ioanna@med.auth.gr (I. Chouvarda), nicmag@med.auth.gr (N. Maglaveras).
0169-2607/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2011.02.008

mainly due to the diversity and heterogeneity of the various storage formats. ECG devices traditionally record ECGs in binary formats which are mainly proprietary, thus not interoperable and comprehensible. Consequently, meaningful ECG data analysis is impeded, also considering that most data formats do not provide the means to store, the equally important, context and content meta-data 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. Retrieving biosignal metadata from other sources (if data are available and procedures are feasible) is often a prohibiting factor for important meta-analysis of the bio-signal [2].

1.1. *Related work in biosignal management*

As effort has been paid by the biomedical community to overcome the shortcomings of the various vendor proprietary formats, open bio-signal standards have been developed. The European standard in electrocardiography domain, SCP-ECG (Standard communication protocol for Computer-assisted electrocardiography), was developed in 1993 and specifies the interchange format and a messaging procedure for ECG cart-to-host communication and for retrieval of SCP-ECG records from the host (to the ECG cart). SCP-ECG standard, as defined by TC251 committee, specifies that information is to be structured in data sections [3]. Besides the actual time-series, and some attached annotations, various types of information regarding the patient and the medical procedure are included, either as mandatory or optional fields [4]. PhysioNet [5] is an Internet resource from the NIH (National Institutes of Health) supplying well-characterized physiological data sets and related open-source software to the biomedical research community. PhysioNet, supports WFDB (Waveform Database) format which is mostly used for ECG data, but provides data format conversions from WFDB to text and EDF plus (European Data Format plus). EDF plus is a 16 bit data format supporting multiple sampling rates and multiple scaling factors, especially suitable for multichannel recordings. It is broadly accepted in the research community for data exchange and many vendors provide import and export filters [6].

Regarding bio-signal archiving and management systems, technological advancements can be seen mainly in two directions. From one side, biosignal management systems have been developed in the context of homecare systems or integration of care and thus have to deal with patient related metadata, as well as possible temporal issues, i.e. multiple signals per patient [7], as well as multimedia and multi-level patient data integration [8]. In parallel, multicentric studies and disease specific research projects have to deal distributed databases and data heterogeneity [9–11], and therefore semantic integration and semantic web techniques [12] have been deployed. In the latter case, where integration of heterogeneous data is required, automated mapping mechanisms need to be developed [13].

In a wider scale and moving beyond specific applications and domains, PhysioNet consists a very good paradigm of a biomedical resource offering access to bio-signals as a public database, however it is not yet structured in a manner that allows for content based search. Mechanisms that would enable the combined access to bio-signals, their features and annotations along with the medical data that describe the context of the bio-signals, would provide a valuable tool, not only for knowledge discovery by the biomedical researcher community, but also for clinical use, in terms of medical evidence and clinical decision support, or even medical error filtering.

1.2. *The proposed work*

In this scope, the proposed work consists a new framework for 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. The proposed work provides a reconciled view of different ECG repositories through the use of ontologies, ECG domain standards and Unified Medical Language System (UMLS) [14].

To demonstrate the potential of the proposed methodology ROISES (Research Oriented Integration System for ECG Signals) framework was developed. Within ROISES development four interesting points were distinguished:

- The definition of a method to integrate diverse biosignal datasources.
- The introduction of a standardized global ontology based on broadly accepted ECG domain standards and UMLS huge variety of source vocabularies and terminologies to virtually aggregate the diverse resources.
- The use of ontologies to support the decoupling of ECG information from the sources.
- The specification of a terminology enhancement method, based on UMLS lexicon and data mining classification algorithms, to support the validity and extensibility of the integration scheme.

The remainder of the paper is organized as follows. Section 2 presents the proposed methodology and ROISES system. Section 2.4 specifically refers to the tools and technologies employed in ROISES. Section 3 elaborates on the results in terms of methodology application and system demonstration. Finally Sections 4 and 5 include the discussion and conclusion part of the paper respectively.

2. Materials and methods

Our approach addresses the problem of content based unified query expression on diverse and heterogeneous bio-signal sources. The approach consists of the following methodological steps:

- The generation of a conceptual model comprising the global schema the source schema and the mapping component.
- The definition of the global ontology, which provides an aggregated view of the underlying sources and allows for query definition with respect to its terminology.
- The specification of the mapping ontology to accommodate the mappings among the global ontology and the source schemas.