Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)

doi:10.3233/978-1-61499-512-8-840

A health analytics semantic ETL service for obesity surveillance

M. Poulymenopoulou[1], D. Papakonstantinou[1], F. Malamateniou[1] and G. Vassilacopoulos[1,2]
[1]*University of Piraeus, Piraeus, Greece*
[2]*New York University, New York, NY, USA*

Abstract. The increasingly large amount of data produced in healthcare (e.g. collected through health information systems such as electronic medical records - EMRs or collected through novel data sources such as personal health records – PHRs, social media, web resources) enable the creation of detailed records about people's health, sentiments and activities (e.g. physical activity, diet, sleep quality) that can be used in the public health area among others. However, despite the transformative potential of big data in public health surveillance there are several challenges in integrating big data. In this paper, the interoperability challenge is tackled and a semantic Extract Transform Load (ETL) service is proposed that seeks to semantically annotate big data to result into valuable data for analysis. This service is considered as part of a health analytics engine on the cloud that interacts with existing healthcare information exchange networks, like the Integrating the Healthcare Enterprise (IHE), PHRs, sensors, mobile applications, and other web resources to retrieve patient health, behavioral and daily activity data. The semantic ETL service aims at semantically integrating big data for use by analytic mechanisms. An illustrative implementation of the service on big data which is potentially relevant to human obesity, enables using appropriate analytic techniques (e.g. machine learning, text mining) that are expected to assist in identifying patterns and contributing factors (e.g. genetic background, social, environmental) for this social phenomenon and, hence, drive health policy changes and promote healthy behaviors where residents live, work, learn, shop and play.

Keywords. Big data, semantics, health analytics, obesity, public health surveillance

Introduction

Obesity is a public health problem that has raised concern worldwide since it is a major cause of co-morbidities, including type II diabetes, cardiovascular diseases and other health problems, which can lead to further morbidity and mortality. The etiology of obesity is multifactorial, involving complex interactions among the genetic background, hormones and different social and environmental factors, such as sedentary lifestyle and unhealthy dietary habits [1]. Recently, the need for a public health approach has been advocated in order to develop population-based strategies for the prevention of excess weight gain, such as the European Childhood Obesity Surveillance Initiative [2]. Obesity prevention studies require the systematic collection, analysis and interpretation of all factors affecting weight gain in order to drive health policy and promote lifestyle, environmental and socioeconomic changes [1,3,4].

Big data analysis can play a key role in both research and intervention activities and provide unique opportunities for monitoring population body weight and obesity. In particular, data from traditional resources like electronic patient records (EPRs) and primary care systems provide clinically rich data while data from other resources like personal health records (PHRs), mobile applications, social media and sensors/devices connected to persons can provide knowledge in patients' health, sentiments and daily activities (e.g. daily exercise) that is extremely important for monitoring routine trends in overweight and obesity in order to understand the progress of the epidemic in population groups [5-8]. However, there are several challenges when integrating big data, including security, interoperability and trustworthiness (data quality). In particular, a factor that hinders big data interoperability is that most healthcare information systems are designed to meet local needs while novel resources produce data with significant noise (low quality) [4,5,9]. In this paper, the semantic data interoperability problem is tackled using appropriate semantics technologies, standard formats, models and terminology systems in order to retrieve, integrate and semantically annotate big data in order to result into valuable data for analysis.

In computing, the Extract-Transform-Load (ETL) process involves extracting data from multiple resources or applications, cleansing this data and loading it into another system like analytics. In order to achieve semantic integration of data, beyond technical integration, there is a need of using a semantic data model using semantic technologies (e.g. ontologies) that enable the creation of semantic mappings among the concepts used in resources and those defined in the semantic data model [5,10,11]. The cloud-based service-oriented paradigm combined with semantic technologies can enable the realization of semantic ETL services. Moreover, healthcare information exchange networks, like the IHE-based networks, that enable patient data sharing in the form of standardized documents (e.g. based on the Clinical Document Architecture - CDA) can provide both a syntactic and a first degree of a semantic basis for the interoperability of clinical concepts between computer systems. However, as regards semi-structured and unstructured data (e.g. from PHRs and social media) that demonstrate significant noise, appropriate analytic and pre-processing techniques (e.g. topic modeling) are required for mapping this data into concepts defined in the semantic data model [11,12].

In this paper, a semantic ETL service on the cloud is presented as part of a health analytics engine that seeks to semantically integrate data from various sources for use by analytics to discover associations and understand patterns and trends within the data that helps providing insights to understanding the factors affecting obesity in order to inform health action and policy. This service uses NoSQL databases for big data storage [13]. An illustrative implementation of the ETL service proposed is also presented in the public health area of human obesity.

The basic motivation for this research stems from our involvement in a project concerning knowledge extraction from big data for factors affecting human obesity. The stringent needs for semantically integrating big data from a variety of sources, for use by data analytics applications, motivated this work and provided some of the background supportive information for the development of the semantic ETL service.

1. Methods

The semantic ETL service proposed may be connected to multiple data sources which are mainly divided into the following categories: a) IHE-based repositories where documents in the form of XML CDA documents with patient information are stored from a variety of sources like hospital information systems (e.g. EMRs) and primary care systems, b) PHRs where documents in the form of Continuity of Care Document (CCD) are exported with patient information, c) sensing technologies like wearable medical devices connected to patient or devices (e.g. smart phones) that transmit patient stream data and d)) other web resources like discussion groups and social networks (e.g. twitter, medical social networks) that store structured to highly unstructured data in the form of messages.

A high level architecture of the semantic ETL service is shown in Figure 1. It is seen that the main modules comprising the service are: data retrieval, data repository and data transformation. In addition to the core ETL process steps (extract, transform, load) the ETL service proposed provides a “store” step among the first two steps. Thus, the modified proves calls for data retrieval, data storage in a schema-less format, data transformation and then data storage in the form of RDF documents. Moreover, in the data transformation step, semantic technologies are used in order to generate semantically annotated data based on the use of a semantic data model. This model provides knowledge of the semantics and structure of data from multiple sources and represents the structure of the selected integration document schema. The use of a semantic data model provides great flexibility when new data sources are added.

The data retrieval module consists of client applications that make calls to the services provided by the data sources and/or by services created in order to connect and retrieve specific patient data (e.g. obese people). For example, these services retrieve a) IHE-based documents with patient information from the IHE repository, b) CCD-based documents with patient information from patient PHR, c) stream data from sensing devices and c) messages from various web sources. Hence, the data retrieved by the data retrieval module is a combination of IHE-based documents, CCD-based documents, stream data and text.

The retrieved data is then stored to NoSQL databases of the data repository module in a schema-less format in order to provide flexibility [13]. For better performance on big data storage and processing, the data repository module supports the use of a combination of document store (e.g. MongoDB) and wide column store (e.g. HBase) NoSQL database types for different types of data. Hence, for each resource according to the data format used, the data velocity (frequency of update) and the data volume, a combination of the above two NoSQL database types is used.

The data transformation module transforms data into RDF documents. This module incorporates services that use a set of analytic techniques for preprocessing data (e.g. normalizing data, removing duplicates). For example, anomaly detection techniques are used to remove abnormal values from sensor data. In addition, topic modeling is used to rank unstructured text (e.g. PHR text, social media messages) according to its relevance to specific topics of interest (e.g. words relative to obesity like food habits). Moreover, this module incorporates an ontology engine, a set of semantic ontologies and a rule base. These ontologies represent in the form of classes (nodes) and relationships (links) the integration document schema, IHE-based and CCD-based document schemas and sensor data context. In addition, a set of ontology rules are defined a) to transform low-level context sensor data into high-level context data (e.g. from accelerometer data range “low-level context” can be derived if a patient is walking or running “high-level context”) and b) to map ontology classes representing concepts of the retrieved data into ontology classes of the integrated document schema. Hence, data of the data storage module is first preprocessed with the use of analytic techniques and then inserted as facts into the ontology under corresponding classes and properties. Then, through ontology reasoning high-level context data is derived and data is transformed into documents compliant with the integration schema. RDF documents are then exported from the ontologies and stored to the data repository module.

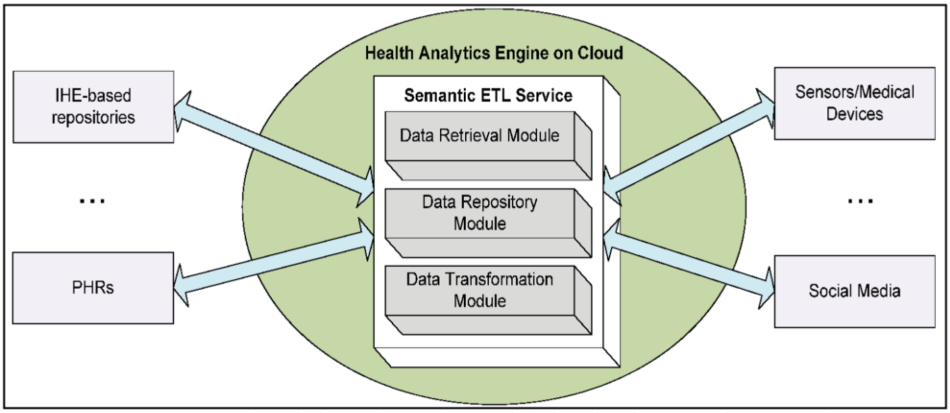

Figure 1. A health analytics engine on the cloud equipped with a semantic ETL service

2. Results

An illustrative implementation of the proposed semantic ETL service is in progress using a laboratory cloud infrastructure, MongoDB and HBase as NoSQL databases, RestFull technology for the implementation of services and the Apache Mahout for its scalable machine learning algorithms. As regards the semantic ontologies, the Protege OWL editor was used for ontology engineering, the Semantic Web Rule Language (SWRL) for ontology rules, the Jess rule engine for rule reasoning and the SPARQL for exporting RDF documents. The semantic ETL service is connected to an IHE-based repository, to a PHR system called PINCLOUD, some sensors/medical devices like accelerometers and pulse wrist sensors and the Facebook.

In this implementation, an integration document schema is created based on a subset of XML elements of the IHE-based and CCD-based documents as well as other elements representing food habits, behaviour and activities that are potentially related to obesity. Therefore, the semantic ontologies created incorporate knowledge in this schema, the IHE-based and CCD-based XML document schemas and sensor data context. Moreover, ontology rules are created that link ontology classes to the classes representing the integrated schema. Currently, services are under development in order to connect to IHE repositories, PHRs, sensors and social media to retrieve patient healthcare data, self-reported data, daily food and exercise habits as well as patients behaviour and sentiments. In particular, the data retrieval services support a) querying data resources according to specified criteria (e.g. selected time window, health problem obesity), b) setting triggers on sensors raw data and c) storing retrieved data to the NoSQL databases. The services of the data transformation module use Apache Mahout functions for pre-processing the data of the data repository module using machine learning algorithms and a set of classes created for inserting retrieved data into the ontologies as facts and resulting into RDF documents after rule reasoning.

3. Concluding remarks and discussion

Semantically integrated data from EMRs, PHRs, sensors and social media can provide increased opportunities for more robust prevention studies on human obesity [1,3,4,8]. However, as quality decisions come from quality data, data pre-processing is critical and considerable work is needed to ensure data consistency and validity across sources, platforms and systems [10,12]. To this end, a special purpose semantic ETL service can provide semantically annotated, rich and meaningful, integrated big data that can be analyzed to produce valuable information and knowledge about the major public health problem of human obesity. In this paper, such a semantic ETL service is presented that seeks to integrate and pre-process data from multiple sources to result into RDF documents. Our intention is to use the integrated data of the proposed service in analytics algorithms in order to provide insights about the factors affecting gaining weight that is expected to drive public health decisions for promoting weight loss and healthy behaviors. Moreover, we plan to test this service with data from different data sources and with data sets and analytics techniques related to different health problems in order to test its performance and results accuracy (e.g. unstructured text mappings to the ontology classes).

References

[1] Frank L., Andersen M., Schmid T., Obesity relationships with community design, physical activity, and time spends in cars, *American Journal of Preventive Medicine* **27:2** (2004).

[2] WHO, European childhood obesity surveillance initiative (COSI), Available from URL http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/activities/monitoring-and-surveillance/who-european-childhood-obesity-surveillance-initiative-cosi. Last accessed 21-10-2014.

[3] Chiolero A., Santschi V., Paccaud F., Public health surveillance with electronic medical records: at risk of surveillance bias and overdiagnosis, *European Journal of Public Health* **23:3** (2013), 350-351.

[4] Ng K., Ghoting A., Steinhubl S. R., Stewart W. F., Malin B., Sun J. PARAMO: A Parallel predictive modeling platform for healthcare analytic research using electronic health records. *Journal of Biomedical Informatics* **48** (2014), 160-170.

[5] Kumar P., Pandeya K. Big data and distributed data mining: An example of future networks,. *International Journal of Advance Research and Innovation.***2** (2013), 36-39.

[6] Lee C., Birch D., Wu C., Silva D., Tsinalis O., Li Y., Yan S., Ghanem M., Guo Y. Building a generic platform for big sensor data application. In IEEE International Congress on Big Data (2013), 94-102.

[7] Guille A., Favre C., Hacid H., Zighed D., SONDY: An open source platform for social dynamics mining and analysis. ACM SIGMOD International Conference on Management of Data. Jun 2013, New York, United States.

[8] Paul M., Dredze M., You are what you tweet: Analyzing twitter for public health. In Proceedings of the fifth International AAAI Conference on Weblogs and Social Media (2011), 265-272.

[9] Lazer D., Kennedy R., King G., Vespignani A., The parable of Google flu: Traps in big data analysis, *Science* **343:6176** (2014), 1203-1205.

[10] Bansal S., Towards a semantic extract-transform-load (ETL) framework for big data integration. In Proceedings of the IEEE International Congress on Big Data (2014), 522-529.

[11] Ferguson M. Architecting a big data platform for analytics. Intelligent Business Strategies, (2012), available from URL http://www.ndm.net/datawarehouse/pdf/Netezza%20-%20Architecting%20A%20Big%20Data%20Platform%20for%20Analytics.pdf

[12] Figo D., Diniz P., Ferreira D., Cardoso J., Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing* **14** (2010), 645-662.

[13] Bonnet L., Laurent A., Sala M., Laurent B., Sicard N. Reduce, you say: What NoSQL can do for data aggregation and BI in large repositories. In 22nd International Workshop on Database and Expert Systems Application, IEEE Computer Society (2011), 483-488.