A content-aware analytics framework for open health data

L. Koumakis[1], H. Kondylakis[1], D.G. Katehakis[1], G. Iatraki[1],
P. Argyropaidas[1], M. Hatzimina[1] and K. Marias[1]

[1] Institute of Computer Science, Foundation for Research and Technology (FORTH), Heraklion, Greece

Abstract—**The vision of personalized medicine has led to an unprecedented demand for acquiring, managing and exploiting health related information, which in turn has led to the development of many e-Health systems and applications. However, despite this increasing trend only a limited set of information is currently being exploited for analysis and this has become a major obstacle towards the advancement of personalized medicine. To this direction, this paper presents the design and implementation of a content aware health data-analytics framework. The framework enables first the seamless integration of the available data and their efficient management through big data management systems and staging environments. Then the integrated information is further anonymized at run-time and accessed by the data analysis algorithms in order to provide appropriate statistical information, feature selection correlation and clustering analysis.**

Keywords—**Data Analysis, Data Mining, Heath data integration, IHE profiles, semantic interoperability.**

I. Introduction

Healthcare is challenged by large amounts of data that is diverse, unstructured and growing exponentially. The heterogeneity and scale of such data (clinical, environmental, lifestyle, etc.) raises the demand for seamless data access along with the availability of powerful, reliable and efficient data analysis operations, tools and services. Obviously, the amount of information available, the heterogeneity of it and the wide range of terminologies/ontologies available to model this information, dictate the identification of a solution able to handle all this data.

Health data analytics provide mechanisms able to identify patterns or trends in data, screen pre-frailty states and provide different views of data for new management plans. Data mining consists of various methods and algorithms, which have been applied to many research areas, and the healthcare domain is not an exception [1, 2]. Understanding and extracting knowledge from healthcare data, formed the need for advanced analytical methodologies that can effectively transform data into meaningful and actionable information [3].

The major challenge in healthcare analytics is not the data mining algorithms, per se, but rather the framework which leverages legions of disparate, structured, and unstructured data [4]. Analytics can provide insides and draw conclusions for the data only if the data source(s) have been appropriately integrated and populated by reliable content.

To this end, we propose an analytics framework over a well-defined mutli-layer approach, which provides efficient management of big data, seamless integration using semantics and standards, secure interaction and anonymization of data for public open access and a modular analytics framework able to incorporate advanced algorithm. The bottom layer can retrieve effectively data from disparate sources and combine those utilizing IHE Profiles [5], the second layer is responsible for the semantic integration and the efficient management of heterogeneous big data, the anonymization and data access, and the third layer provides a user-friendly analytics portal. The proposed framework has been adopted by the iManageCancer European project [6] as an open access tool to any researcher for analysing anonymized, health related data. The framework was empirically evaluated by experts using artificial but realistic data that exist in real medical databases such as patients' admission details, demographics, laboratory exams, medications, wellbeing data from smart phones and smart watches, etc.

This paper describes a framework for data analytics over health related big data sources. It presents the system's architecture in Section 2 by means of describing (i) the heterogeneous health sources integration by means of the IHE technical framework, (ii) the staging big data environment and the semantic layer, both feeding (iii) the Data Analysis Layer. Section 3 provides preliminary results from the iManageCancer project and section 4 concludes.

II. System architecture

Analytical services try to go much further than traditional statistics by examining the raw data and then attempting to hypothesize relationships within the data. As shown in Fig. 1 on the top, data analysis and data mining is an iterative approach, which combines data from the semantic layer and the big data staging environment, pre-processes the data, performs the analysis and provides the results for visualization based on the data distillation model. The loop closes with the interaction of the end user who can refine the results and continue with a drill-down analysis to extract knowledge from patient cohorts with specific criteria.

The proposed architecture consists of three layers. The data analytics layer, the semantic layer and the heterogeneous health sources integration layer. In the next sections, we analyze in detail each one of the aforementioned layers.

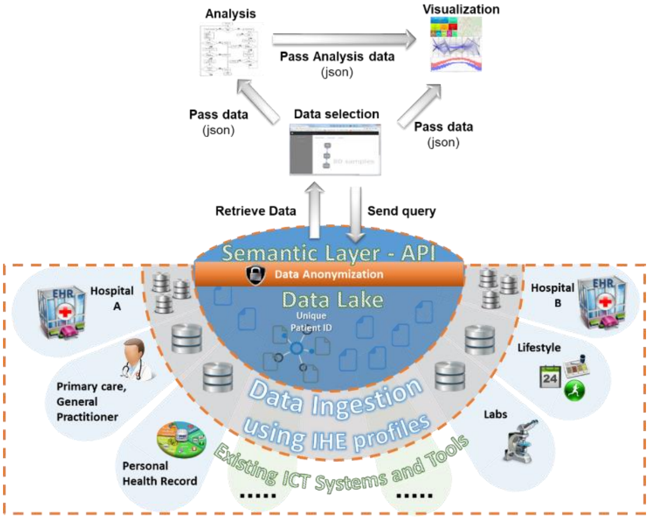

Fig. 1: The reference architecture of the proposed analytics framework

A. Heterogeneous health sources integration using IHE

IHE profiles have the potential to support the sharing of health information in a secure, reliable and incremental manner across the different points of care, through authorized and validated interaction with existing systems and tools. This requires the existence of a commonly agreed interoperability framework to be in place, including (amongst other) detailed conformance statements for each domain under consideration. Implementations of software must be in accordance with the specifications described by IHE Profiles [5], such as those for Cross Document Sharing (XDS), Patient Identifier Cross Referencing (PIXv3), Patient Demographics Query (PDQv3), Cross Community Access (XCA), and Cross Community Patient Discovery (XCPD) for peer-to-peer querying and retrieve with other communities.

Cross-organization health data sharing translates into complex security policies that need to be uniformly managed and enforced. New complex requirements include for example the capability of dealing with data-binding concepts such as ‘purpose of use’ and ‘conditions on use’ [REF[3]].

The IHE Quality, Research and Public Health (QRPH) domain [10] addresses the information exchange and electronic health record content standards that are necessary for the sharing of information relevant to quality improvement in patient care, clinical research and public health monitoring. IHE QRPH addresses the infrastructure and content necessary to share information relevant to quality improvement, improve the liaison between the primary care system and clinical research and provide population base health surveillance, which are all reliant on the secondary use of data gathered in clinical care. Some examples of relevant IHE QRPH profiles indicatively include the Clinical Research Process Content [11] and the Research Matching [12]**.**

Having in our disposal the IHE profiles, we can rely on a health information ecosystem, with well-defined interoperability standards for every health related sources such as Electronic Health Records, Personal Health Records, lifestyle monitoring, laboratory results, etc.

B. Semantic Layer and Application Programming Interface

The semantic layer consists of two main sub-layers, i.e. the big data lake and the integrated information layer. On top of these layers APIs use various data services to provide access to the available information exploiting anonymization services according to the security/ethical requirements. Bellow we explain in detail each one of those components.

Data Lake Layer

Presented architecture's big data lake layer corresponds to IHE ingestion layer. In the data lake, various information exists in its proprietary form and is further processed and cleaned as a first step of data management. Various databases are staged in this layer such as PostgreSQL storing the PHR patient information, MySQL for storing the recommendations to the patients and the corpus to be recommended, Cassandra DBs for staging big data available (e.g. activity monitoring data, sensor data etc.) and other DBs for storing the pushed information from external sources.

Integrated Information Layer

Selected information out of the data lake layer is mapped to a modular ontology, the IMC Semantic Core Ontology, developed as part of the iManageCancer project[4]. Then, data are semantically uplifted (through an ETL process), and stored in a Virtuoso Triple Store. A benefit of the approach is that we can recreate from scratch the resulting triples at any time.

[3] D. W. Chadwick and S. F. Lievens, “Enforcing ‘sticky’ security policies throughout a distributed application,” MidSec ’08 Proc. 2008 Work. Middlew. Secur., pp. 1–6, 2008.

[4] Haridimos Kondylakis, Anca Bucur, Feng Dong, Chiara Renzi, Andrea Manfrinati, Norbert Graf, Stefan Hoffman, Lefteris Koumakis, Gabriella Pravettoni, Kostas Marias, Manolis Tsiknakis and Stephan Kiefer, iManageCancer: Developing a platform for Empowering patients and strengthening self-management in cancer diseases, 30th IEEE International Symposium on Computer-Based Medical Systems - IEEE CBMS

However, for reasons of efficiency the data integration engine periodically transforms only the newly inserted information by checking the data timestamps. In this process summarization tools [13] allow the quick exploration and understanding of the available information enabling subsequent query formulation.

Data Access Services (APIs)

Both the integrated information and the staged information at the data lake can be queried using appropriate Data Access Services through the appropriate APIs. The APIs transform the user request to the appropriate query language (CQL, SQL or SPARQL) and forward the query to the appropriate source. The analytics framework usually queries the integrated information since we would like to analyze the linked information between the sources.

C. Data Analysis Layer

The objective of the analytical framework is to extract information from the diverse health data sources and transform it into an understandable structure for better knowledge and further use. The platform is modular enough and allow any data-mining algorithm to be directly embedded in the whole workflow. As we can see from the high-level architecture (Fig. 1 on the top), the components of the analytical framework are the query builder, analysis and visualization.

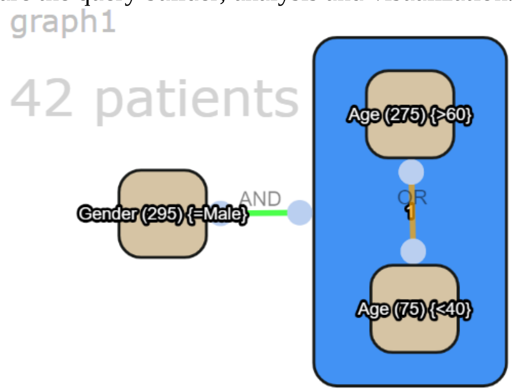

Fig. 2: Query builder example

Query Builder

Query builder is the place where the end user poses the research question and pull the anonymized data from the data lake. Since the end users can be also non-IT experts the graphical interface intended to be simple but yet powerful enough for complex queries. The user actually draws an SQL-like query with his/her preconditions and selects the attributes that would like to retrieve data from. The implementation is based on graphs and the user has to create/draw a graph where each node is a feature with specific conditions (e.g. the user wants to view data for all the patients with age over 18), while the edges represent the logical condition between the features. The query builder provides the possibility to the user to create more complicated queries using groups e.g. *(age under 40 OR age over 60) AND gender male* as shown in .

Visualization

Query builder provides the capability to the user to view more statistics of the generated query at any time and update/modify the query accordingly. The results of each query can be viewed in a graphical way, using various charts, enabling further exploration and enhancement. Each chart can be used as a filter and give instant feedback. The graphical view of the query results from an example query is shown in Fig. 3: Features with numeric values such as Age are visualized as bars charts while the nominal features such as Gender are visualized as pies. The total number of patients is shown to the right of the viewer area. All the charts can be used as single filters or multiple filters (using the logical AND operation for more than one filters).

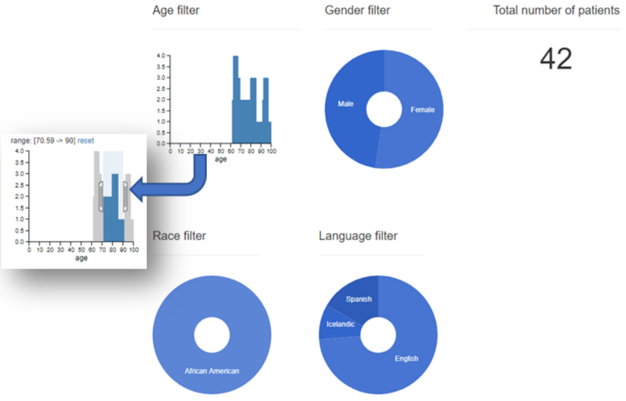

Fig. 3: Visualization

Data Analysis

Data mining consists of various methods and algorithms, which have been applied to many research areas and the healthcare domain is not an exception. Main objective of the analytics framework is to hide the complexity of data mining from the end user.

Handling of the diverse and large amount of data is supported by feature elimination algorithms, typically used in the demanding domain of bioinformatics [14, 15] while clustering algorithms provide similarity matching to cases/patients. The reduction of the feature set and the selection of the most relevant features could help to cope with highly dimensional data (e.g. lifestyle data from smart devices), reduce computational cost, and improve classification performance. The framework uses the principal variables [16] in order to select a subset of variables that contain, in some sense, as much information as possible and propose the most informative variables of a cohort to the end user.

One of the most important questions in data analysis is to find the "similar" cases/records in our data. Cluster analysis has been widely used in patient orientated management strategies and identify discrete groups of patients with specific combinations of comorbid conditions [17]. The analytics framework uses the K-Means algorithm [18], one of the well-known unsupervised learning algorithms, which clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares.

III. PRELIMINARY EVALUATION AND DISCUSSION

The proposed data analysis framework has been adopted by the iManageCancer European project. The iManageCancer project aims to support chronic cancer treatment via a cancer disease self-management platform focusing on the wellbeing [19] and patient empowerment [20].

As such the analytics framework has been linked with the iManageCancer data sources using the semantic and interoperability layers. During the implementation of the aforementioned analytics platform, the iManageCancer pilots were in the process of preparation. Therefore, no real data were available. For that reason, we generated artificial but realistic data for testing and development of the framework and the algorithms. The framework, using the artificially generated data, was empirically evaluated by four experts (two physicians, one research nurse and one data analyst) As it was expected the end users focused mainly on the visualization of data and analysis rather that the analytical algorithms. They identified the whole framework as a highly added-value instrument, easy to be understood and used. The query builder was the only part of the analysis workflow that the users were not familiar but after a while, using the trial and error method, all the users managed to create simple and complex queries. The results of a thorough evaluation, with real-datasets will be published in a follow-up after the completion of the pilots.

IV. Conclusions

Nowadays, medicine combines data collected over time about an individual's genetics, environment, and lifestyle and focuses on the integrated diagnosis, treatment and prevention of disease in individual patients [21]. While the goal is clear, the path to such advances has been fraught with roadblocks mainly in the data management and data integration areas.

We propose a multi-layer framework architecture and we believe that consolidating healthcare data into comprehensive and coherent assets with an analytics frontend on top will aid in the precision medicine area. The architecture is able to combine heterogeneous healthcare sources by exploiting IHE profiles, integrate efficiently big data using semantics and provide anonymized healthcare data over a modular analytics framework. The proposed architecture/ framework has been adopted by the iManageCancer project but it could also be used out of the project's context.

Acknowledgment

This work has been supported by the iManageCancer H2020 EU programme under grant agreement No 643529.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Yoo I, Alafaireet P, Marinov M, et al (2012) Data mining in healthcare and biomedicine: A survey of the literature. J Med Syst 36:2431–2448. doi: 10.1007/s10916-011-9710-5
2. Potamias G, Koumakis L, Moustakis V (2005) Mining XML Clinical Data: the HealthObs System. Ingénierie des systèmes d'information 10:59–79.
3. Reddy CK, Aggarwal CC (2015) Healthcare data analytics. CRC Press
4. Belle A, Thiagarajan R, Soroushmehr SMR, et al (2015) Big Data Analytics in Healthcare. Hindawi Publ Corp 2015:1–16. doi: 10.1155/2015/370194
5. IHE. https://www.ihe.net.
6. iManageCancer. http://imanagecancer.eu.
7. XDS. http://wiki.ihe.net/index.php/Cross-Enterprise_Document_Sharing.
8. PQGv3. http://wiki.ihe.net/index.php/Patient_Demographics_Query_HL7_v3.
9. XCA. http://wiki.ihe.net/index.php/Cross-Community_Access.
10. QRPH. https://www.ihe.net/Quality_Research_and_Public_Health/.
11. Clinical Research Process Content. http://wiki.ihe.net/index.php/Clinical_Research_Process_Content.
12. Research Matching. http://wiki.ihe.net/index.php/Research_Matching.
13. Pappas A, Troullinou G, Roussakis G, et al (2017) Exploring Importance Measures for Summarizing RDF/S KBs. In: Blomqvist E, Maynard D, Gangemi A, et al (eds) Semant. Web 14th Int. Conf. ESWC 2017, Portoro{ž}, Slov. May 28 -- June 1, 2017, Proceedings, Part I. Springer International Publishing, Cham, pp 387–403
14. Potamias G, Koumakis L, Moustakis V (2004) Gene Selection via Discretized Gene-Expression Profiles and Greedy Feature-Elimination. In: Methods Appl. Artif. Intell. Third Helenic Conf. AI, {SETN} 2004, Samos, Greece, May 5-8, 2004, Proc. pp 256–266
15. Koumakis L, Moustakis V, Zervakis M, et al (2012) Coupling regulatory networks and microarays: Revealing molecular regulations of breast cancer treatment responses. In: Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). pp 239–246
16. McCabe GP (1984) Principal Variables. Technometrics 26:137–144.

doi: 10.2307/1268108

17. Newcomer SR, Steiner JF, Bayliss EA (2011) Identifying subgroups of complex patients with cluster analysis. Am J Manag Care.
18. MacQueen JB (1967) Kmeans Some Methods for classification and Analysis of Multivariate Observations. 5th Berkeley Symp Math Stat Probab 1967 1:281–297. doi: citeulike-article-id:6083430
19. Kondylakis H, Kazantzaki E, Koumakis L, et al (2014) Development of interactive empowerment services in support of personalised medicine. Ecancermedicalscience 8:
20. Kondylakis H, Koumakis L, Kazantzaki E, et al (2015) Patient Empowerment through Personal Medical Recommendations. In: Stud. Health Technol. Inform. p 1117
21. Huang BE, Mulyasasmita W, Rajagopal G (2016) The path from big data to precision medicine. Expert Rev Precis Med Drug Dev 1:129–143. doi: 10.1080/23808993.2016.1157686