QUALITATIVE EVALUATION OF DISTRIBUTED CLINICAL SYSTEMS SUPPORTING RESEARCH TEAMS WORKING ON LARGE-SCALE DATA

TOMASZ DZIUBICH

Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
Narutowicza 11/12, 80-233 Gdansk, Poland

(received: 11 August 2017; revised: 13 September 2017;
accepted: 20 September 2017; published online: 6 October 2017)

Abstract: In this paper, five contemporary scalable systems to support medical research teams are presented. Their functionalities extend from heterogeneous unstructured data acquisition through large-scale data storing, to on-the-fly analyzing by using robust methods. Such kinds of systems can be useful in the development of new medical procedures and recommendation rules for decision support systems. A short description of each of them is provided. Further, a set of the most important features is selected, and a comparison based-on it is performed. The need for high performance computing is emphasized. A general discussion how to improve the existing solutions or develop new ones in the future is also presented.

Keywords: medical informatics, system architecture, data acquisition, data analysis, medical procedures, recommendation rules

DOI: https://doi.org/10.17466/tq2017/21.3/k

1. Introduction

Researchers today are more focused on creating an effective logical pattern derived from archival medical data in order to understand the behavior of the body with respect to the disease. For this purpose, previous medical data has become an important asset in data analysis and the prediction process for future diseases, symptoms and treatment. However, there is a key issue with the effective utilization and quality assurance of the data. To reach these goals suitable software and hardware platforms for medical data acquisition, storing, processing, sharing and tagging, have to be provided [1]. The authors have indicated that a wide range of clinical data is stored in electronic health records (EHRs) and can be potentially used to conduct medical research. However, EHRs are dominated by unstructured narrative data that is not available for research or quality improvement efforts. Clinicians often have relevant well-defined and structured databases. Both these sets usually disjunct.

The purpose of having quality information systems is to produce and use quality information. To provide safe care to patients, clinical staff must use highly reliable hospital information systems (HIS) to provide the required information. Information needs within hospitals are currently expanding, and the majority of hospitals now have huge servers, storing and processing electronic medical records (EMRs), and many terminals allowing clinical staff to input or browse EMRs. Information networks connect servers and terminals. Information is vital for the safe care of patients and, hospitals are driven by information, but it is insufficient for medical researchers in most cases [2].

All kinds of medical data are included in disparate categories of clinical information systems software (CIS) for healthcare on the institutional level: HIS (hospital information system), LIS (laboratory information system), RIS (radiology information system) and PACS (picture archiving and communication system). Besides those mentioned above, there are personal health systems, and finally systems supporting evidence collection for research teams, especially in the epidemiology field. Not all such systems incorporate valuable, complete and high quality data for medical research.

One of the issues addressed by scientists is bioimage analysis and annotation, which are necessary to diagnose and monitor *e.g.* an oncology disease, extracting relevant information useful to make a correct diagnosis, and to define an adequate treatment facilitating the exchange of information between care centers [3].

Due to the growing popularity of diagnostic methods based on large-scale data, in particular deep neural networks, we still face the problem of creating IT platforms supporting medical research teams. The requirements for such software that we can list include:

- accessing, acquiring and storing heterogeneous data from existing EMR and medical devices (including PACS) in a distributed manner;
- sufficient computational power for on-the-fly analysis;
- integration with medical devices which do not support the DICOM format;
- dealing with many structured and unstructured data sources;
- easy creation of research scenarios (research workflows) to test hypotheses;
- tightly-coupled cooperation with other teams in the same field/specialization.

These requirements arise from a common practice in the process of the creation of new medical procedures which is shown in Figure 1. This process assumes that the medical team uses sophisticated computer software.

The most popular model of medical research is still based on measurements and their statistical analysis. In the first step, a single researcher or a team gather data. Next, the collected data is complemented with additional information

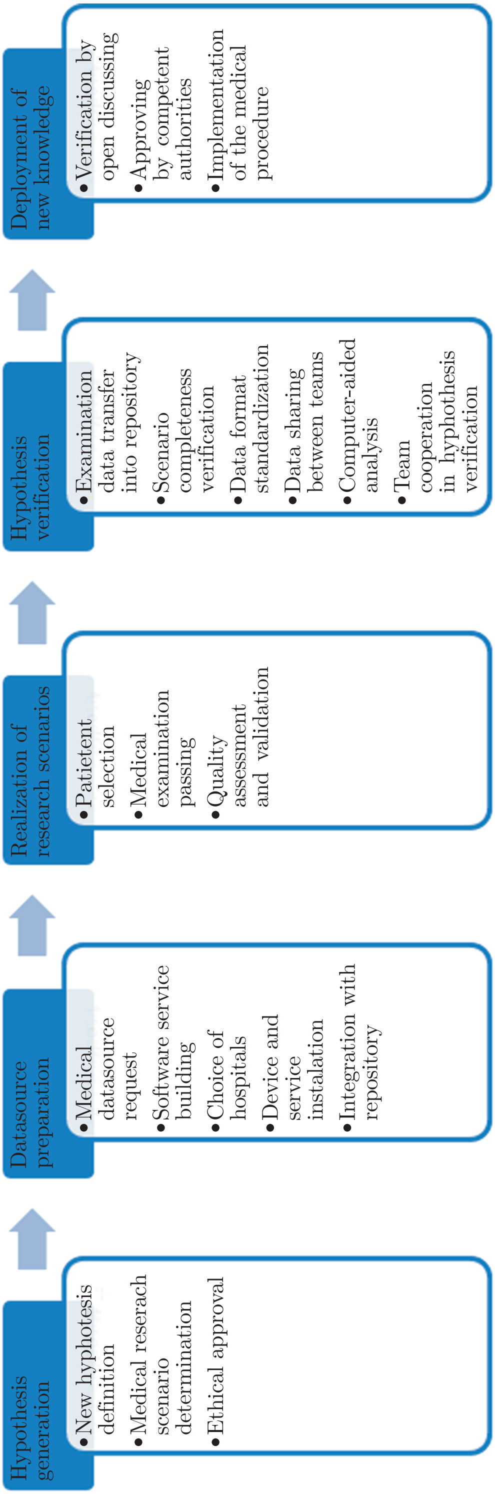
Hypothesis generation
• New hyphotesis definition
• Medical reserach scenario determination
• Ethical approval
Datasource preparation
• Medical datasource request
• Software service building
• Choice of hospitals
• Device and service instalation
• Integration with repository
Realization of research scenarios
• Patietent selection
• Medical examination passing
• Quality assessment and validation
Hypothesis verification
• Examination data transfer into repository
• Scenario completeness verification
• Data format standardization
• Data sharing between teams
• Computer-aided analysis
• Team cooperation in hyphothesis verification
Deployment of new knowledge
• Verification by open discussing
• Approving by competent authorities
• Implementation of the medical procedure

Figure 1. Workflow of new medical procedure creation

fundamental to further analysis. This is the hardest stage, due to ensuring the completeness of treatment results. Both methods are contingent on access to a big tagged set of measurements (annotation), images, videos and discharge (or histologic) diagnoses. Statistical methods are primarily based on surveys.

To gather the described data, researchers use sophisticated modules accessing HIS (*e.g.* Medical Data Repository and Medical Analysis System for CLININET [4]) which come in handy with collecting sets of biomedical data (images, laboratory test results, patient records, epidemiological analyses etc.) and creating the workflows (pipelines) used to process such data. On the other hand, the collected data could be shared with other scientists or teams. One of the most famous frameworks for medical collaboration is Taverna [5]. It is an environment for preparation, visualization and execution of bioinformatic workflows that integrates various tools available on web resources. It enables easy integration with Openoffice, Statistica. Many systems based on the framework: myExperiment, BioBomy can be found in the papers. Taverna is based on the web services technology. Unfortunately searching, acquisition and running appropriate services is quite difficult for medical research teams.

In [6] the authors have proposed a medical 3D image reconstruction system which uses the processing and analysis of a series of 2D CT images to convert to a 3D model, which provides 2D and 3D images, a medical image processing unit, surgical marker space coordinates, a system control function, etc. Through the help of the navigation system, the doctor can operate images with various operations, such as zooming, rotating, measuring, and marking. They have developed a test environment and carried out scene tests many times.

Gomez *et al.* [7] proposed a modular and flexible multimodal monitoring platform, aimed to support medical research that requires processing biological variables acquired at the patient bed-head. It is compact and lightweight which makes it especially suited for intensive care environments, and can be connected to a broad range of standard monitoring equipment. Not only mean values, but also complete waveforms are registered, which greatly improve the data analysis possibilities. A powerful software package has been developed, to acquire, manage and process patient data; with a modular architecture to easily include new features and processing blocks. Up to now, biosignals of about 30 patients have been monitored, registering more than 100 hours of examinations. The cerebral hemodynamics toolbox is being used at several projects, and some of the results have already been presented.

Another interesting solution is described in [8]. To deal with the congestion problems when different types of business flows compete for a limited bandwidth in the ESB, a real-time ESB model is proposed. The model uses a priority-driven bandwidth allocation strategy according to different business type bandwidth requirements. A simulation environment is built based on the Regional Medical Information Exchange Platform. Simulation results show that this method can guarantee a bandwidth of a high-priority business, and meet the needs of a real-time business. On the other hand, traffic scheduling is used to avoid the decline of QoS caused by congestion, and improve the overall system throughput.

In this paper, we discuss five different systems which predict notable aspects of the patient's health, and guide effective disease management to the patient. Two of them were designed, implemented and deployed by a team conducted by the author of this paper. According to our knowledge, it is a comparison of the newest medical IT systems supporting the collaboration of researchers. The highlighted features are a modern software architecture, readiness to work on big medical data, and sharing it between interested persons.

The paper is organized as follows. In Section 2 a brief description of five chosen systems is presented. These are: FIU, CSDC, RPDMBS, IPMed and ERS. Thereafter, the distinguishing characteristics are selected, and a comparison of these systems is made. Finally, we discuss the general rules and requirements which have to be met in modern IT systems for medical research teams.

2. Review of existing systems for distributed data acquisition and processing

For professional information technology research, the above mentioned systems are valid; however, their practical implementation becomes difficult due to the complexity of understanding for each application in an emergency and fast-paced environment. This also results in wasting more time in understanding and frequent use of these systems, rather than using more time with patients, due to lesser knowledge availability to doctors about EHR systems and their benefits.

In further analysis, we resign from the above mentioned systems and, choose others which, according to us, have a more flexible software architecture, and operate on larger data. The systems had to lie on the intersection of three fields of interest: distributed systems, health informatics, clinical medicine. The next idea which motivated us to choose representative systems was modernity. Systems could not be older than five years. Another topic was the ease of expansion and migration to other medical areas (ability of generalization). Thus, the preferred architecture was service-oriented or similar. The last issue was the supporting collaborative work of research teams. This requirement is more difficult, due to assurance of security and privacy.

Finally, five systems were selected for comparison. FIU [9], CSDC [10], ERS2012 [11], IPMed [12] and RPDMBS [13] are different medical web-based platforms that try to address the medical researcher's problem for data collection and understanding. These systems are very intricate for generating a complete and comprehensive solution in the general domain. Thus, the authors were showing a practical application in a selected field. The interfaces of each application vary from one another. Patient data and past history are also derived by different methods and also the resulting visualizations are different in each application and tool that are a bit complex for physicians to understand. The general architecture of such systems is shown in Figure 2.

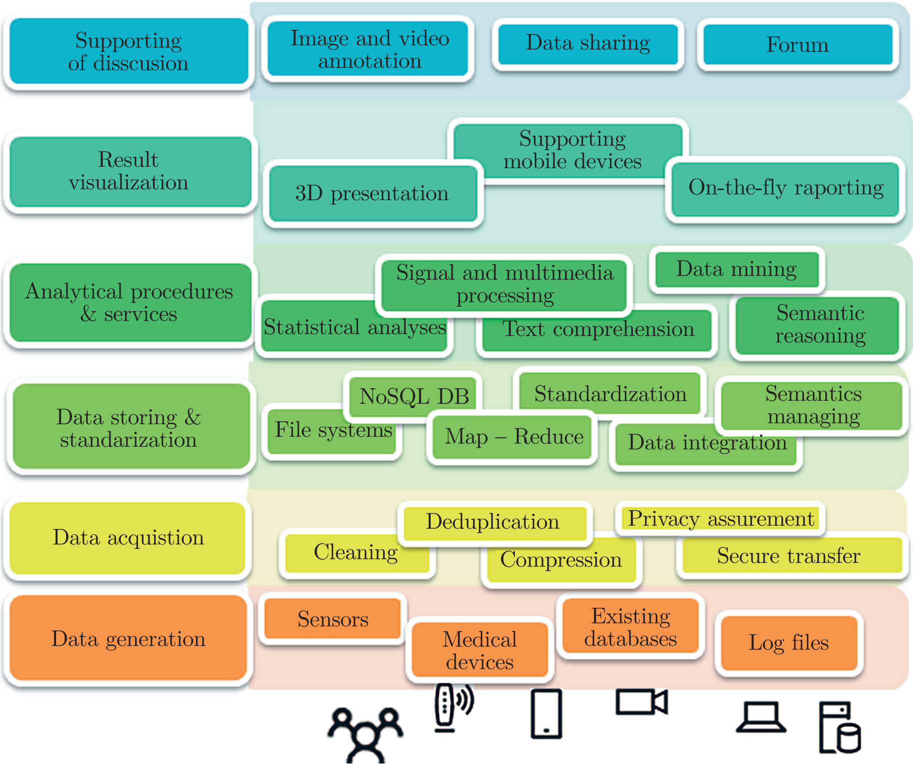

Figure 2. Multi-layered general architecture of modern medical system working on large-scale data

2.1. FIU – Platform for Alzheimer's disease

In [9] the authors have described a web platform for data acquisition and analysis for Alzheimer's disease that was created at the Florida International University, Miami (FIU). The main objective of the FIU is to develop methods to automatically process MRIs by employing FreeSurfer (FS) – a worldwide known, robust tool used for MRI brain segmentation – as a web service [14]. Nevertheless, processing of a single MRI is very time-consuming (from several hours, to a day, depending on the processing pipeline).

The FIU also enables automatic acquisition and processing of data for neurological studies with a focus on Alzheimer's Disease (AD). The FIU platform consists of:

- A data upload module that provides forms and web services to store volumetric and surface area measures, derived from the MRI to a MySQL database. Uploaded MRI files are sent to the FS web service through the processing server;
- A Web Interface and Web Services for classification of AD, utilizing Support Vector Machines (SVM). In order to conduct a recommendation process, the user has to input other features – gender, age, education, and the technical

parameters of the device. By using an SVM classifier four output classes of diagnosis are supported. The output results are sent to the user through e-mail;

- A Web service for segmentation – the MRI is processed with FreeSurfer (FS); the FS output is also made available to the Web Interface user.

The FIU design is scalable and can be easily adapted to other neurological disorders.

From a technological point of view, the platform is based on Drupal CMS, MySQL and several WS servers for data processing. The hardware configuration mentioned by the authors is capable of processing an MRI within 10–12 hours. At full capacity (running one MRI per core), eight MRIs can be processed concurrently, with an average of 16 MRIs per 24 hours.

2.2. A Nationwide Screening Platform for Stroke Control and Prevention (CSDC)

Another distributed platform for medical data collecting and processing is CSDC [10]. It is aimed at providing information needed for decision making for stroke control and prevention, clinical evaluation and research, public health service in China. The screening data is acquired in a two-fold way: from the high-risk population and from EMRs of individual patients with stroke. Structured questionnaires are used to collect information on risk factors, diagnosis history, treatment, and sociodemographic characteristics as well. Among the gathered data we can enumerate: past medical history, mode of life, family history, medication history, physical examination, and laboratory examination. The important issue is to take the patients' survival on board (follow-up data). The initial mode of stroke onset, initial and follow-up diagnostic and neurologic evaluations, and therapy have been also collected for individuals who were stroke patients. These selected individuals should be followed-up four times (3 months, 6 months, 1 year and 2 years after first reported). The platform has a built-in quality check procedure. All data collection forms follow the standards established by national and provincial specialists.

The CSDC platform consists of:

- China Stroke Data Integrator (CSDI) – a configurable adaptor/interceptor mechanism to build a stroke database for base hospitals and to integrate EMR to the CSDC central database. The proposed mechanism employs the ICD-10 terminology and a web portal which was deployed at base hospitals, to adaptor configuration, data validation and integration which will be sent to the CSDC central database.
- Large data analysis environment – the core of the system includes interactive and flow analysis services; stroke domain specific analysis algorithms are planned to be studied and implemented as software components; the whole of the component is complemented by data integration tools – Sqoop, Spark, HDFS, Hive YARN.

- The analysis environment provides interfaces to R library, Python library and development API for a developer to create many own applications.

2.3. Endoscopy Recommended System (ERS2012)

ERS2012 is inextricably bounded up with the KASKADA platform (Supercomputer Platform for Context Analysis of Data Streams in Identification of Specified Objects or Hazardous Events) which is placed on the Galera supercomputer in the Academic Computer Centre TASK [11]. The primary aim of KASKADA is to share different kinds of software services, particularly those demanding high computational power. Except being a powerful execution environment for time-consuming algorithms, KASKADA also provides a universal external interface in the form of automatically created Web services, enabling launching algorithms from remote locations, *e.g.* from the doctor's office. KASKADA is also a framework facilitating construction of stream algorithms.

An example of using the KASKADA platform is the ERS2012 system (a part of which is the MedEye application), supporting physicians in the gastrointestinal endoscopy image analysis (Wireless Capsule Endoscopy, WCE) and diagnosis. ERS2012 enables medical doctors to fast scan an endoscopic video. The suspected frames on the video stream are highlighted, together with the most probable diagnosis.

The ERS2012 system consists of the following components:

- client applications (MedEye) – graphical user interface (for browsers and as a WPF application);
- database of medical evidence – repository in which anonymized patient data is stored (annotations and selected areas of lesions);
- flat set of endoscopic videos – a decentralized file system for archiving of pictures and videos, connected to the database;
- disease detection algorithms – a set of robust algorithms (*e.g.* bleeding, gastropathy, polyp detection) formed as a web service.

2.4. Software platform for supporting medical research teams (IPMed)

The IPMed platform was designed as a thorough solution for medical researchers seeking to gather and analyze large amounts of diverse medical data [12]. It has been successfully deployed in 5 independent hospital wards, where it enables gathering extensive data of stroke cases, including impedance cardiography (ICG) examination records. The general functionality of the platform can be outlined as follows: medical data gathering and storing, enabling data analysis supporting tools, and supporting a remote teamwork of researchers.

The IPMed system is based on a microservices architecture, where they are communicating with each other through the network using the REST (Representational State Transfer) architecture style.

The central point of the system is the main repository which gathers anonymous data that it receives from local hospital data servers. Each hospital gathers data from medical equipment and physicians. The data stored in these hospitals is encrypted, and decryption requires an authorized user to authenticate. On the other side of the system, there is a cooperation webserver that serves a website for cooperation of physicians. Webservices supporting the search for the most likely evidence in the impedance cardiography are placed on a webserver.

The IPMed system consists of the following components:

- Cooperation web server – it provides a facility of cooperation of physicians from different hospitals/medical centers by giving them access to the data stored in the main repository in one common and uniform data format. The platform contains a recommender system based on an expert system that classifies patients by their hemodynamic profile. It can be extended by additional services to support diagnosis in other fields.
- Main repository – it is a well-secured webserver which is only accessible by applications that have a trusted encryption key that is used for encrypting communication. All the data kept on this server is completely anonymous (anonymization of this data is done on local hospital data servers before this data ever leaves the source hospital). Depending on the data source type, flat or relational database engines are used.
- Local hospital data servers and data source devices – devices that gather medical data are connected to local hospital data servers (being a part of the IPMed platform) and upload data to them. The authors have employed Manatec PhysioFlow and Medis NICCOMO cardiographs as data source devices. Each hospital has its own local data server which contains patients' medical data that is stored in an encrypted database. This is the only part of the system that stores sensitive data. The key used for encrypting the database is also encrypted, but it is done using user keys that are not stored on the same server. The local data server connects to the main repository in fixed intervals in order to send the data.
- User devices in hospitals – data stored on local hospital data servers is accessed via user devices like tablets (Android, iOS or other devices with a web browser). The devices have preinstalled IPMed client applications which connect to the local hospital data server via RESTful web services. These kinds of devices are used to present the recommendation (the most likely class of hemodynamic profile).

These data and analysis results would become important resources to study morbidities and order of risk factors in different catalogues in Poland. Investigation of the data may reveal the stroke status and trends in Poland by utilizing the IPMed platform. And furthermore the data can also be used for clinical research and administrative statistics.

2.5. Research Platform for Building Medical Diagnostic Services (RPDMBS)

The RPDMBS is a research platform whose purpose is to minimize the limitations of its technical use, as it makes maximally wide medical applications possible. The main contribution is semi-automatic creating of decision supporting systems on the base of medical knowledge retrieval. The methods of diagnostic rules development, and the poly-procedural approach have passed evaluation tests while the tissue dysplasia syndrome is being researched (hypermobility syndrome, HS) [13]. We cannot give any detailed results as no more papers about this platform are available in public.

In the RPDMBS a multilayer architecture is used which contains the following levels:

- data storing – one interesting thing is the using of the Continuity of Care Record standard (CCR) as an alternative to the Clinical Document Architecture (CDA) of Health Level 7 (HL7) standard for data storing. CCR is designed for storing private medical data. CCR allows describing the medical data on the object to the full extent, and is widely spread: such medical services as Google Health and Microsoft HealthVault are based on it. This layer is supplied with authentication and personalization methods;
- data adaptation – this layer is responsible for integrating the research platform database into exterior medical software systems, *e.g.* clinical information systems (CIS) of healthcare institutions. The translation format task is done by software agents. No specific adapter/format is mentioned;
- applied services – the layer provides a set of tools which are useful for researchers. We can, *inter alia*, list: analytical systems, data collection services, a diagnostic rules generation module and a composition subsystem of medical diagnostic services.

The system allows recognizing statistical relationships between symptoms and diagnoses. It also generates a software service that allows running computer-assisted diagnosis on the basis of the data on the disease collected by a researcher.

The authors have presented the results of such a system, incorporated in the tissue dysplasia syndrome research. They have compared four statistical methods: frequency, frequency with correlations, entropy, and the likelihood ratio. The quality of the diagnostics of the suggested approaches is about 88.5% of correct recognized objects on the training sample and about 83.5% on the test sample. The poly-procedural approach consolidates diagnosis of several methods and increases the diagnostic reliability by 10–15% compared with each separate method.

3. Features selection and system comparison

Feature extraction and representation is a fundamental stage in the comparison process, so it is important to carefully choose the right features for any system. The feature selection process for comparison of a specialized general-purpose medical platform is quite hard. There is no recommendation for any standard. Vendors of healthcare software often use attributes such as flexible, easy-to-use, accessible, streamlined, and multidisciplinary to promote their products (based on comparison matrix technique). However, this is often at odds with the principles of data security, which talk about privacy and confidentiality. Most of the decision supporting medical systems are the content-based image retrieval systems; and explore low-level image features such as color, texture, shape, motion, and so on, because they can be computed automatically. In our case they are completely useless with regard to the possible descriptive nature of data.

On the other hand, for example, taking into account the characteristics of the research data by the data source, we can divide systems by the data quality, availability of clinical data elements, follow-up data or representativeness of the study cohort. Unfortunately, not all of those features are described in papers. Therefore, we had to consider a suitable abstraction level of the selected features.

The major research objective of our work is to develop a general approach to build a dedicated system supporting medical diagnostics. Based on our experiences in this area we can point out two essential steps: 1. Provide support for medical research teams for acquisition of valuable data and storing it in a big data repository; 2. Develop efficient software tools which will improve big medical data analysis.

To achieve that aim we used the comparative analysis method (in a normative manner). The first step in this method is selection of alternative cases, then an arrangement of criteria and comparative analysis/study in aspect of dominant features. Selected advantages are essential for the development of further systems.

Based on our knowledge related to that subject, we chose general aspects of medical IT systems and then a small set of attributes, which occur in the bulk of papers. The attribute selection process was conducted from four points of view: provided/offered functions, the applied system architecture, technological aspects, and supported research methods. The selected attributes and their detailed description are presented in Table 1.

In the considered system the following three processing models are analyzed:

- Generic processing model: This model is addressing general application problems and the model used is MapReduce. MapReduce is a really plain and potent model which enables automatic distribution and parallelization of computation spread across big clusters of machines. MapReduce is based on two user defined functions, and called Map and Reduce. The Map function is grouping all intermediate values related to the intermediate key, and sends them to the Reduce function. The Reduce function gets the data and merges it to obtain a smaller number amount of values. Putting SQL on top of MapReduce is an effective solution to lower the learning curve for traditional programmers who are experienced in SQL.
- Graph processing model: Many applications can be described in terms of objects that are connected one to another using graph models. Computational jobs are specified as directed graphs. Every vertex is a user-specified value. Source vertices are connected to directed edges, and every edge has a modifiable value and destination vertex id. When the application is initialized, the program is executed as a chain of iterations, separated with

Table 1. Selected attributes to compare the analyzed systems

Aspects	Attribute name	Description
Provided/ required system functionalities	Way of gathering data	Manual, automated or semi-automated data inputting, kind of data source (clinical data, anonymized clinical data for medical education, anonymized epidemiological and public health data, personal data), form of records (text, scan, raw image, movie), data quality (completeness, reliability)
	Data storing	Used data volume – (GigaBytes vs. constantly updated TeraBytes/PetaBytes), its structure (homogenous vs. heterogeneous, structured vs. unstructured, semi-structured), standardized data (*e.g.* DICOM, XML, HL7 CDA, FHIR) velocity (generated rate – per day vs. per hour or more rapid)
	Visualization	Processing site (hospital, computing centre, cloud), technique (geometric, graph-based, pixel-oriented), displaying (mobile, web-based, specialized)
	Data analysis	Descriptive, predictive or prescriptive analytics, quantity of built-in method, ability of own research method placement/ using (procedure, flow), time of execution analyses (batch or on-the-fly), type of analysis (data visualization, statistical analysis, data mining)
	Primary purpose	Epidemiological studies, decision making support, recommendation, build own methods to analysis
Architecture characteristics	Reference model	Client/server, multi-tier, broker, web services, microservices
	Programming model for data	Map-reduce (generic batch processing), streaming (pipeline), graph processing
	Infrastructure	Private or public cloud (IaaS), dedicated cluster, supercomputer, stand-alone
	Data management	Distributed file system (plain), noSQL database, SQL database
	Interoperability	Possibility of data flow between different systems, available protocol converters, used communication standards
Technologies and solutions	Implementation platforms	Programming language (Java, C#, Python), front-end technology (JS, HTML 5, ML), additional (CUDA, OpenCL)
	User interface	Graphical or textual, web-form based, non-interactive or interactive, customized workflow composition
	Data sources	Quantity and types of applied devices/field of interest to feed a system
	License	Type of used license: open-source or commercial
User oriented propositions	Utilized approaches	Questionnaires/surveys/statistics algorithm-based decision making self-improving of decision making
	Data quality assurance	Use of preprocessing techniques and dictionaries, cross-validation, on-line validation
	Number of users/teams	Estimated number of users or teams which operate on relevant systems
	Reachability	Global, regional, local
	Practical use	Academic, applied in practice

synchronization barriers until the algorithm calculates output. Each iteration executes the user-specified function parallel in the vertex. The vertex can modify its state and outgoing edges, it can send and receive messages to and from other vertices, or change the graph topology. Each vertex can deactivate itself, when all vertices are stopped at the same time the application terminates. The output of such application is a directed graph which is isomorphic to its input.

- Stream processing model:. Keyed tuples in the data stream are considered as an event and directed to a specified processing element (PE). Processing elements are acyclic graphs which control the processing of events with specific keys and sending results. Processing nodes are hosts to processing elements, and they are waiting for events and then redirecting them to the processing element container, which is activating right PEs in the right order.

Taking into account data processing approaches, the most important roles are played by the following functions:

- Data visualization: is closely related to information graphics and information visualization. The goal of data visualization is to communicate information clearly and effectively through graphical means. In general, charts and maps help people understand information easily and quickly. However, as the data volume grows to the level of big data, traditional spreadsheets cannot handle the enormous volume of data. Visualization for big data has become an active research area because it can assist in algorithm design, software development, and customer engagement.
- Statistical analysis: is based on statistical theory which is a branch of applied mathematics. Within statistical theory, randomness and uncertainty are modeled by the probability theory. Statistical analysis can serve two purposes for large data sets: description and inference. Descriptive statistical analysis can summarize or describe the collection of data, whereas inferential statistical analysis can be used to draw inferences about the process. More complex multivariate statistical analysis uses analytical techniques such as regression, factor analysis, clustering, and discriminant analysis.
- Data mining: is the computational process of discovering patterns in large data sets. Various data mining algorithms have been developed in the artificial intelligence, machine learning, pattern recognition, statistics, and database communities. During the 2006 IEEE International Conference on Data Mining (ICDM), the ten most influential data mining algorithms were identified based on rigorous selection. In the ranked order, these algorithms are C4.5, *k*-means, SVM (Support Vector Machine), a priori, EM (Expectation Maximization), PageRank, AdaBoost, kNN, Naive Bayes, and CART. These ten algorithms cover classification, clustering, regression, statistical learning, association analysis and link mining, which are all among the most important topics in the research on data mining. In

addition, other advanced algorithms, such as neural network and genetic algorithms, are useful for data mining in different applications.

Moreover, Blackett [15] has divided data analytics into three categories depending on the analysis depth: descriptive, predictive and prescriptive analytics.

- Descriptive analytics: uses historical data to represent what happened. For example, regression can be used to find trends in data. Visualization techniques are applied to portrait data in a telling way. Data modeling is used to store, collect and reduce data in an effective fashion. Such analysis is commonly used in business intelligence.
- Predictive analytics: aims to predict trends. For instance, such models use statistical methods like linear or logistic regression to determine and predict trends or future outcomes.
- Prescriptive analytics: resolves decision making. Simulations are used to research complicated systems and obtain knowledge about their behavior, and can help find optimizations or optimal solutions for given parameters.

Based on system descriptions and the meaning of attributes shown in Table 1 we estimated them in Table 2.

As is shown in Table 2 the most promising system is CSDC. This system has been applied in China and is still developed. Unfortunately, its source code is closed, hence detailed analysis is impossible. The most valuable advantage is using 'true' big data. On the other hand, the provided user interface is very poor. ERS2012 and IPMed have much more sophisticated ones.

4. Conclusions

Medical doctors across the world have started to explore various technological solutions to enhance medical procedures in a manner that complements the existing analytic methods by tapping the potential of the large scale data. This paper surveys diverse aspects of service-based research technologies in healthcare, and presents various systems/platforms that support access to facilitate medical data transmission, gathering, storing and processing.

In the development process of modern medical research systems we can observe a tendency to the acquisition of large scale data which is valuable from a scientific point of view, and on the other hand, to share it with the community. Such medical data mainly has the following three aspects of characteristics: polymorphism, incompleteness and redundancy. Medical information may contain pure data (such as sign parameters and test results), signals (*e.g.*, ICG, etc.), images (such as MRI, CT and other medical imaging equipment), texts (such as the patient record identity and symptom description, detection and diagnosis results of text expression), as well as voice and video information. The multi pattern feature is the most prominent feature of the data, and meanwhile, the coexistence of the multiple attribute patterns increases the difficulty of medical data mining. The medical database is a huge data resource, every day there is a large amount of the same, or some of the same, information stored in it. This redundancy of

Table 2. Evaluation of satisfaction level regarding selected features

Attribute	FIU	CSDC	ERS	IPMed	RPDMBS
Way of gathering data	L	H	M	H	M
Data storing	L	H	H	M	L
Visualization	M	L	H	H	M
Data analysis	Prescriptive analytics	Predictive analytics	Prescriptive analytics	Prescriptive analytics	Descriptive analytics
Primary purpose	Decision making support	Epidemiological studies + decision support	Recommendation	Recommendation	Developing one's own methods to analysis
Reference model	Web services	Web services	Web services	Microservices	Web services
Programming model for data	Batch	Map-reduce	Streaming	Streaming	Graph processing
Infrastructure	Cluster	Private cloud	Supercomputer	Private cloud	Stand-alone
Data management	L	H	H	M	H
Interoperability	L	H	L	H	H
Implementation platforms	Python, Drupal PHP, HTML	Java, Python, HTML	C++, CUDA, WPF, HTML	Java, JS, HTML	Java, HTML
User interface	L	L	H	H	L
Data sources	L	H	M	H	H
License	Open-source	Commercial	Open-source	Open-source	ND
Utilized approaches	Statistics	Questionnaires/surveys/statistics	Algorithm-based decision making	Self-improving of decision making	Self-improving of decision making
Data quality assurance	M	H	H	H	M
Number of users/teams	1	1000	1	6	1
Reachability	Local	Global	Regional	Regional	Regional
Practical use	No	Yes	Yes	Yes	Yes

Level of satisfaction (L – low, M – medium, H – high) medical information makes the medical data mining different to other general data mining, and makes the medical data mining become very specific. Despite that, in most cases, the lack of valuable data can occur, and the expression and the record of much medical information has the characteristics of uncertainty and fuzziness. In addition to the above, there are no typical BigData systems with regard to the trade-off between quality and amount of data.

These conditions require that the following clues are provided in the development process improving the functionality of developed/offered systems, and the use of high-performance computing services, to achieve:

- Increasing flexibility and simplicity which has a huge impact on the system interoperability. Using a service-oriented architecture seems like an indispensable aspect.
- Achieving distributed on-line processing which enables speeding-up medical treatment procedures and faster diagnoses.
- Including self-learning attributes. Due to the growing number of observed parameters which mostly exceeds human perception, new approaches to analyzing are needed, especially those based-on machine learning techniques.
- Increasing the set of collected data and its analysis. It is possible to accomplish it by improving the functionality, security assurance and sharing data between medical teams.
- A user friendly interface is essential to increase the popularity of systems. In the presented systems the utilized interfaces are simple web forms or batches without advanced graphical interfaces which are often used in medical workstations. The new approaches based-on graphical composition of service workflow are necessary.

In the future, we plan to develop the ERS2012 and IPMed systems to reach the full set of the above described requirements, and we will continue to study how to realize service scheduling in distributed ESB clusters; to achieve high performance and scalable real-time information integration platforms.

Acknowledgements

This research was funded by grants from the National Centre for Research and Development (PBS2/A3/17/2013, Internet platform for data integration and collaboration of medical research teams for the stroke treatment centers).

References

[1] Katzan I L, Rudick R A 2012 *www.ScienceTranslationalMedicine.org* **4** (16228)

[2] Richardson I, Reid L, O'Leary P 2016 *Healthcare Systems Quality: Development and Use*, Proc. of IEEE/ACM International Workshop on Software Engineering in Healthcare Systems 50 doi: doi:10.1109/SEHS.2016.01

[3] Vizza P, Guzzi P H, Veltri P, Cascini G L, Curia R, Sisca L 2016 *GIDAC: A prototype for bioimages annotation and clinical data integration*, 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 1028 doi: doi: 10.1109/BIBM.2016.7822663

[4] *Clininet documentation* [Online] Available at: https://www.cgm.com/pl/prodcuts___solutions/medical_colleges/solutions_for_universities.pl.jsp [accessed 5.06.2017]

[5] Wolstencroft K, *et al.* 2013 *The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud*, *Nucleic Acids Research* **41** (W1) W557 doi: doi:10.1093/nar/gkt328

[6] Chen Y Sun P 2014 *The research and practice of medical image 3D reconstruction platform*, Proceedings IEEE International Conference on Security, Pattern Analysis, and Cybernetics (SPAC) 305 doi: doi:10.1109/SPAC.2014.6982704

[7] Gomez H, Camacho J, Yelicich B, Moraes L, Biestro A Puppo C 2010 *Development of a multimodal monitoring platform for medical research*, 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology 2358 doi: doi:10.1109/IEMBS.2010.5627936

[8] Xia C, Song S 2011 *Research on real-time ESB and its application in regional medical information exchange platform*, 4th International Conference on Biomedical Engineering and Informatics (BMEI) 1933 doi: doi:10.1109/BMEI.2011.6098735

[9] Lizarraga G, Cabrerizo M, Duara R, Rojas N, Adjouadi M, Loewenstein D 2016 *A Web Platform for data acquisition and analysis for Alzheimer's disease*, SoutheastCon 1 doi: doi:10.1109/SECON.2016.7506730

[10] Yu J, Mao H, Li M, Ye D, Zhao D 2016 *CSDC – A nationwide screening platform for stroke control and prevention in China*, 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2974 doi: doi:10.1109/EMBC.2016.7591354

[11] Jędrzejewski M, Brzeski A, Blokus A, Cychnerski J, Dziubich T 2012 *Real-Time Gastrointestinal Tract Video Analysis on a Cluster Supercomputer*, Springer 55

[12] Dorożyński P, Brzeski A, Cychnerski J, Dziubich T 2016 *Towards Healthcare Cloud Computing, Information Systems Architecture and Technology*, Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 (ed. Świątek J.), Springer 87 doi: doi:10.1007/978-3-319-28564-1_8

[13] Apanasik Y, Shabalina I, Kuznetsova L 2013 *The research platform for building medical diagnostic services*, 14th Conference of Open Innovation Association FRUCT 9 doi: doi:10.1109/FRUCT.2013.6737939

[14] Reuter M, Rosas H D, Fischl B 2010 *Highly Accurate Inverse Consistent Registration: A Robust Approach*, *Neuroimage* **53** (4) 1181

[15] Blackett G *Analytics Network – O. R. Analytics* [Online] Available at: http://www.theorsociety.com/Pages/SpecialInterest/AnalyticsNetwork_analytics.aspx [accessed 5.06.2017]