Jurnal Pencitraan Digital (2019) 32:870-879
https://doi.org/10.1007/s10278-019-00184-5

Kerangka ETL untuk Intelijen Bisnis Real-Time atas Medis Repositori Pencitraan

Tiago Marques Godinho[1] & Rui Lebre[1,2] & João Rafael Almeida[1,2] & Carlos Costa[1]

Dipublikasikan secara online: 31 Januari 2019
Masyarakat untuk Informatika Pencitraan dalam Kedokteran 2019

Abstrak

Dalam dekade terakhir, jumlah studi pencitraan medis dan metadata terkait telah meningkat pesat. Meskipun sebagian besar digunakan untuk mendukung diagnosis dan perawatan medis, banyak inisiatif baru-baru ini mengklaim penggunaan studi pencitraan medis dalam skenario penelitian klinis tetapi juga untuk meningkatkan praktik bisnis institusi medis. Namun, produksi terus menerus dari studi pencitraan medis ditambah dengan sejumlah besar data terkait, membuat analisis real-time dari repositori pencitraan medis menjadi sulit menggunakan alat dan metodologi konvensional. Arsip tersebut tidak hanya berisi data gambar itu sendiri tetapi juga berbagai metadata berharga yang menggambarkan semua pemangku kepentingan yang terlibat dalam pemeriksaan. Eksplorasi teknologi tersebut akan meningkatkan efisiensi dan kualitas praktik medis. Di pusat-pusat utama, ini mewakili skenario data besar di mana Business Intelligence (BI) dan Data Analytics (DA) jarang terjadi dan diimplementasikan melalui pendekatan data warehousing. Artikel ini mengusulkan kerangka kerja Ekstrak, Transformasi, Muat (ETL) untuk repositori pencitraan medis yang dapat memberi makan, secara real-time, aplikasi BI (Business Intelligence) yang dikembangkan. Solusinya dirancang untuk menyediakan lingkungan yang diperlukan untuk memimpin penelitian di atas repositori institusional langsung tanpa meminta pembuatan gudang data. Ini menampilkan dasbor yang dapat diperluas dengan bagan dan laporan yang dapat disesuaikan, dengan antarmuka berbasis web intuitif yang memberdayakan penggunaan teknik penambangan data baru, yaitu, berbagai alat pembersihan data, filter, dan fungsi pengelompokan. Karena itu,

Kata kunci PACS. Intelijen Bisnis. DIKOM. Analisis Data. awan. Data besar

pengantar

Saat ini, repositori pencitraan medis berisi berbagai metadata berharga yang menggambarkan secara menyeluruh semua pemangku kepentingan yang terlibat dalam praktik pencitraan medis. Meskipun sebagian besar digunakan untuk mendukung diagnosis dan pengobatan medis, banyak inisiatif baru-baru ini mengklaim kegunaan studi pencitraan medis dalam skenario penelitian klinis dan dalam peningkatan praktik bisnis institusi medis.

Paradigma repositori pencitraan medis saat ini sangat cocok dengan definisi data besar [1]. Produksi terus menerus dari volume data yang sangat besar, sifatnya yang heterogen, dan meningkatnya jumlah pemeriksaan yang dilakukan membuat analisis repositori pencitraan medis menjadi sangat sulit untuk alat konvensional. Selain itu, tren baru arsitektur Arsip Gambar dan Sistem Komunikasi (PACS) terdistribusi yang memungkinkan untuk menggabungkan beberapa institusi dalam arsip PACS yang sama di cloud [2] mempromosikan pembuatan kumpulan data yang besar dan lebih berguna. Oleh karena itu, teknik DA dan BI yang diterapkan pada skenario ini berpotensi meningkatkan efisiensi dan kualitas praktik medis.

Artikel ini mengusulkan kerangka kerja ETL untuk repositori pencitraan medis yang memberi makan, secara real-time, platform BI yang berorientasi pada praktik dan penelitian pencitraan medis. Solusinya dapat mengindeks sumber data yang berbeda dan bertujuan untuk menyediakan lingkungan yang diperlukan untuk melakukan penelitian di atas repositori institusional langsung. Ini memanfaatkan semua metadata yang disimpan di dalamnya

* Rui Lebre
ruilebre@ua.pt

Tiago Marques Godinho
tmgodinho@ua.pt

João Rafael Almeida
joao.rafael.almeida@ua.pt

Carlos Costa
carlos.costa@ua.pt

[1] Universitas Aveiro, DETI/IEETA, Kampus Universitário de Santiago, Aveiro, Portugal

[2] Departemen Teknologi Informasi dan Komunikasi, Universitas A Coruña, A Coruña, Spanyol

repositori tanpa memerlukan gudang data, model data yang telah ditentukan sebelumnya, atau memaksakan aliran data yang kaku. Sistem yang dikembangkan memanfaatkan Dicoogle'fitur penambangan data [3] untuk mengekstrak data dari PACS produksi dan menyediakan serangkaian teknik eksplorasi dan alat visualisasi untuk pemahaman mendalam tentang kumpulan data yang berfungsi dan ekstraksi informasi berharga. Selain itu, desainnya memfasilitasi penggunaan alat analitik tanpa memerlukan keterampilan pemrograman pengguna yang biasa digunakan di platform lain (misalnya, Python dan R). Ini menyediakan antarmuka berbasis Web intuitif yang memberdayakan penggunaan teknik penambangan data baru, yaitu, berbagai alat pembersihan data, filter, dan fungsi pengelompokan. Selain itu, ia memiliki dasbor yang dapat diperluas dengan bagan dan laporan yang dapat disesuaikan.

Latar Belakang

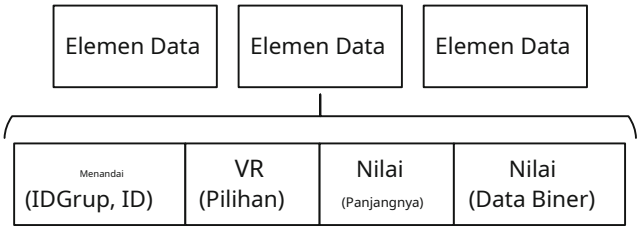
DICOM

Proliferasi PACS dimungkinkan sebagian besar karena pengembangan Digital Imaging and Communications in Medicine (DICOM), standar untuk penanganan data pencitraan medis. PACS bertanggung jawab untuk menyediakan layanan penyimpanan, transmisi, dan bahkan pencetakan, antara lain, untuk memungkinkan konektivitas, kompatibilitas, dan pengoptimalan alur kerja antara peralatan pencitraan medis yang berbeda [4].

Standar DICOM tidak hanya mendukung data piksel yang mendefinisikan citra medis tetapi juga berbagai informasi metadata yang terkait dengan semua pemangku kepentingan yang terlibat dalam praktik klinis, seperti pasien, prosedur, peralatan, data terkait staf, atau laporan terstruktur. Data relatif terhadap pemangku kepentingan ini disampaikan oleh elemen data DICOM yang menyusun objek atau file DICOM.

Elemen data DICOM dikodekan menggunakan struktur Tag-LengthValue (TLV). Bidang tag mengidentifikasi elemen data dan mencakup dua subbidang: pengidentifikasi grup dan pengidentifikasi elemen dalam grup, keduanya dikodekan menggunakan nilai 16-bit yang tidak ditandatangani. Elemen data DICOM dikelompokkan berdasarkan hubungannya dengan entitas dunia nyata, yaitu, Entitas Informasi (IE) yang mewakili, misalnya, pasien (0 × 0010), penelitian (0 × 0008), dan seri (0 × 0020). Oleh karena itu, elemen yang menyimpan informasi terkait pasien dicakup dalam kelompok pasien (0 × 0010) dan seterusnya. Sebagai contoh, pasien'tag nama s diwakili oleh (0 × 0010, 0 × 0010). Selain tag, elemen data DICOM juga menyertakan panjang bidang dalam byte. Terakhir, bidang nilai menyimpan elemen aktual'datanya. Sebuah ilustrasi sederhana dari elemen DICOMdata disajikan pada Gambar.1.

Objek DICOM adalah istilah umum untuk menggambarkan file DICOM, yang dapat berupa gambar, laporan struktur, dan lain-lain. Informasi yang terlampir dalam objek DICOM sangat beragam. Ada elemen data untuk mewakili nama,



Gambar 1 data DICOM struktur elemen

langkah-langkah, tanggal, antara lain. Oleh karena itu, untuk mengekspresikan semua tipe data ini, pengkodean bidang nilai berubah sesuai dengan elemen'tipe s. Bagian 5 dari standar DICOM [5] mendefinisikan 27 format pengkodean yang berbeda untuk bidang nilai. Ini adalah tipe data paling dasar dalam standar dan disebut Representasi Nilai (VR). Elemen data dapat mencakup sub-grup elemen, yang memiliki SQ (Sequence) VR. Ini menciptakan struktur dokumen hierarkis yang mirip dengan banyak model data kontemporer. Struktur ini diilustrasikan pada Gambar.2. elemen data's VR dapat dideklarasikan secara eksplisit dengan memasukkan bidang VR ke dalam elemen's struktur TLV, sehingga mengubahnya menjadi Tag-VR-Length-Value, atau sebagai alternatif, dapat disimpulkan secara implisit dengan membaca kamus standar DICOM [6] yang berisi hampir 2000 entri [7]. Elemen-elemen ini mencakup dengan sangat baik persyaratan umum lingkungan pencitraan medis. Meskipun demikian, standar ini dapat diperluas dan elemen data pribadi dapat ditambahkan oleh produsen untuk mendukung fitur terbaru mereka. Dengan demikian, kamus pribadi dapat memperpanjang default yang disediakan oleh standar. Dengan demikian, standar DICOM memiliki kemampuan untuk mengikuti solusi mutakhir, aspek mendasar untuk prevalensinya di lapangan.

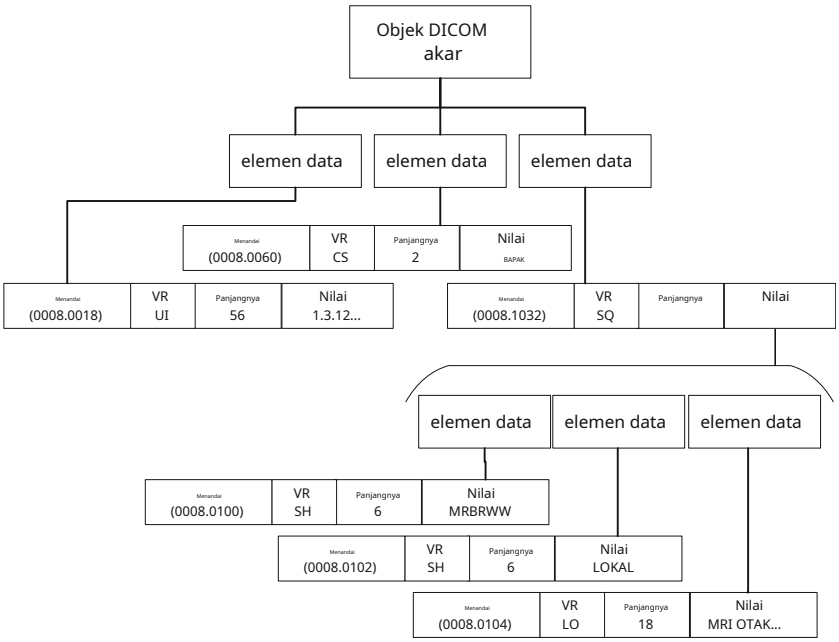
Objek DICOM dapat mewakili beberapa artefak pencitraan medis, seperti gambar dari berbagai modalitas atau laporan terstruktur (SR). Objek-objek ini terdiri dari beberapa modul yang terkait dengan IE, yang mewakili pemangku kepentingan kata nyata seperti pasien dan studi. Mereka mencakup beberapa elemen data; misalnya, modul pasien terdiri dari pasien's nama, jenis kelamin, ulang tahun, di antara atribut lainnya.

DICOM juga mendefinisikan modul mana yang harus disertakan untuk setiap kelas objek DICOM, sesuai dengan definisi model informasi (IOD) spesifiknya [7]. IOD adalah kumpulan modul yang menggambarkan setiap objek. Mereka menentukan informasi mana yang harus disertakan dalam file DICOM, bersama dengan jenisnya.

DICOM mencakup konsep instans yang merupakan templat IOD yang diisi dengan data dunia nyata dan diidentifikasi oleh pengidentifikasi unik (UID). Setiap objek DICOM's instance UID harus unik.

Akhirnya, standar menyediakan Model Informasi DICOM (DIM) yang mengikuti organisasi hierarkis dalam gambar seri studi pasien. Pendekatan ini menyerupai organisasi dunia nyata karena pasien dapat melakukan beberapa studi, setiap studi mungkin memiliki beberapa rangkaian banyak modalitas dan setiap prosedur dapat menghasilkan koleksi besar

Gambar 2. Ilustrasi objek bersarang di DICOM



contoh gambar. Karena gambar DICOM biasanya disebarkan oleh banyak file untuk kenyamanan. Keterkaitan antara entitas ini dicapai dengan menggunakan instance UID. Akibatnya, DIM dapat direkonstruksi bahkan ketika gambar tersebar di beberapa repositori. Atribut UID umumnya dinamai menurut nama Pemangku Kepentingan; namun, ada pengecualian seperti SOPInstanceUID (ServiceObjectPair) yang merupakan UID dari seluruh objek DICOM.

Pencitraan Intelijen Bisnis Business

Analisis isi arsip PACS telah ditunjukkan untuk menarik hasil positif untuk banyak upaya penelitian, yaitu, pengawasan dosis radiasi [8], analisis kinerja proses praktik bisnis institusional [9, 10], efektivitas biaya prosedur diagnostik [11], di antara banyak lainnya [11-14].

Namun demikian, kompleksitas data pencitraan medis telah meningkat pesat [15] karena volume dan heterogenitas data yang dihasilkan oleh praktik medis juga meningkat pesat. Oleh karena itu, peneliti harus mengandalkan perangkat Teknologi Informasi (TI) untuk dapat melakukan tugas analitis mereka. Namun, PACS tradisional tidak mengizinkan penjelajahan metadata pencitraan untuk ekstraksi pengetahuan yang relevan. Ini mengarah pada penggunaan aplikasi pihak ketiga untuk melakukan analisis data, termasuk solusi eksklusif yang hanya bekerja dengan PACS tertentu.

BI terdiri dari pipa yang mengintegrasikan serangkaian tugas. Ketika digabungkan, ini bertanggung jawab untuk memperoleh, mengubah, dan menerjemahkan data mentah menjadi informasi yang berguna untuk meningkatkan praktik bisnis [16]. Proses ini merangkum banyak kemampuan seperti pelaporan, dasbor, dan penambangan data [17].

Saat ini, data yang dihasilkan oleh laboratorium pencitraan medis sangat heterogen dan tidak konsisten. Meskipun semua peralatan menerapkan standar DICOM yang sama, mereka mungkin menggunakan konfigurasi yang berbeda. Ini menghasilkan data yang tidak teratur, yang terjadi, misalnya, ketika dua peralatan berbeda melaporkan nilai yang sama tetapi menggunakan metrik yang berbeda [18]. Selain itu, interaksi antara staf teknis dan peralatan merupakan faktor inkonsistensi lainnya. Untuk alasan ini, menerapkan prosedur analitis untuk data yang tidak konsisten dapat menghasilkan hasil yang tidak dapat diandalkan. Jadi, salah satu langkah terpenting dan krusial dalam keseluruhan proses BI adalah Pembersihan Data (DC) tahap. Ini memastikan keandalan data di setiap repositori, dengan mendeteksi dan mengoreksi catatan yang tidak akurat [19]. DC adalah proses yang kompleks (Gbr.3), yang dapat dibagi lagi menjadi serangkaian operasi berulang. Pertama, langkah Data Auditing, di mana dataset kerja dianalisis untuk menentukan anomali mana yang dikandungnya. Selanjutnya, langkah Spesifikasi Alur Kerja mendefinisikan serangkaian operasi yang diperlukan untuk memperbaiki anomali yang diidentifikasi sebelumnya (atau, dalam beberapa kasus kritis, mengecualikannya). Setelah ini, operasi dijalankan di langkah Eksekusi Alur Kerja. Akhirnya, operasi yang diterapkan divalidasi pada langkah Pasca Pemrosesan dan Pengendalian. Masing-masing langkah ini dapat dilakukan secara manual, terawasi atau tidak terawasi sesuai dengan tingkat input sistem yang dibutuhkan pengguna.

Dicoogle

Dicoogle adalah arsip PACS open source yang memiliki sistem berbasis dokumen pengindeksan dan pengambilan yang dapat diperluas

Gambar 3 Proses pembersihan data

[20]. Ini memberikan perluasan fungsionalitas baru yang mudah melalui pengembangan dan penggunaan plug-in menggunakan Software Development Kit (SDK) yang tersedia. Pembuangan SDK memungkinkan pengembang untuk memperluas fungsionalitas tanpa membuat perubahan pada inti sistem.

Dicoogle dirancang untuk mendukung ekstraksi informasi, pengindeksan, dan penyimpanan metadata yang terkandung dalam file DICOM tanpa persyaratan konfigurasi ulang khusus [21].

Arsitektur yang dapat diperpanjang ini memungkinkan penggunaan Dicoogle dalam penelitian dan industri perawatan kesehatan dan sebagai alat pendidikan [2, 22] karena kebutuhan untuk meningkatkan, memantau, dan mengukur sistem pencitraan medis bersama dengan ekstraksi pengetahuan dari citra medis termasuk indikator kualitas perawatan kesehatan adalah penting.

Selain itu, Dicoogle digunakan sebagai platform dukungan untuk penambangan data DICOM [22]. Alat penambangan data DICOM terbukti berharga untuk mengumpulkan data yang relevan untuk peningkatan layanan kesehatan profesional dengan mengidentifikasi faktor-faktor yang dapat berkontribusi pada defisit kualitas [23].

Pekerjaan yang berhubungan

Ada beberapa referensi dalam literatur yang berkaitan dengan penggunaan sistem DA dan BI yang diterapkan pada repositori pencitraan medis. Nagi dkk. [12] mengembangkan alat yang mengimplementasikan tumpukan BI yang lengkap. Dimulai dengan mengumpulkan data dari institusi's, yaitu, metadata DICOM dari PACS dan data Tingkat Kesehatan 7 (HL7) dari Sistem Informasi Radiologi (RIS), antara lain. Data diekstraksi secara berkala dan disimpan dalam database MySQL yang memberi makan dasbor. Ini adalah alat grafis yang mencakup jenis bagan yang paling relevan, seperti histogram dan bagan gelembung. Kapanpun database'entri ditambahkan atau diperbarui, bagan yang sesuai secara otomatis ditampilkan untuk mencerminkan perubahan ini, serta detail laporan yang dipilih saat ini.

Kalman dkk. [20] mengembangkan kerangka kerja yang menyediakan serangkaian fungsi serupa. Namun, ada beberapa perbedaan penting. Misalnya, kerangka kerja kedua memisahkan header metadata DICOM dari gambar's data piksel, menyimpan yang pertama di repositori terpisah. Selain itu, antarmuka pengguna adalah baris perintah berdasarkan Structured Query Language (SQL). Ini adalah bahasa yang kuat dan fleksibel untuk pakar database tetapi lebih sulit untuk melakukan analisis data untuk pengguna yang kurang berpengalaman, seperti peneliti radiologi.

Dua kerangka kerja sebelumnya menggunakan repositori terpisah; yaitu, mereka tidak bekerja secara langsung atas arsip PACS institusional. Keuntungannya adalah dapat digunakan untuk tujuan statistik tanpa risiko mengganggu alur kerja klinis reguler. Namun, mereka akan selalu bekerja dengan snapshot arsip yang sudah ketinggalan zaman's konten. Dalam sudut pandang kami, batasan utama adalah bahwa mereka terikat pada model basis data relasional yang ketat. Ini berarti bahwa peneliti tidak dapat memperoleh pengetahuan dari bidang metadata yang sebelumnya tidak tercakup dalam skema database, tidak memenuhi persyaratan sebagian besar upaya penelitian dengan cara ini.

Wang dkk. [8] mengembangkan solusi database untuk mengontrol pasien's paparan radiasi dengan menganalisis metadata gambar DICOM. Ini difokuskan pada deteksi dosis radiasi tidak teratur oleh beberapa filter dan dapat mengumpulkan informasi di tingkat penelitian, pasien, dan institusi. Ini termasuk pelaporan, kemampuan peringatan, dan antarmuka Web untuk berinteraksi dengan sistem. Modul yang paling menarik, dalam konteks artikel ini dan membandingkan dengan alat sebelumnya, adalah Basis Pengetahuan yang mampu menyatukan data vendor yang berbeda dengan menerapkan atribut yang diukur untuk menggunakan unit yang sama, yang didefinisikan secara statis. Namun, kendala utama dari solusi ini adalah tidak dapat digunakan dalam konteks pencitraan lain.

Di [18], penulis menyajikan Gudang Data DICOM untuk penambangan data sewenang-wenang. Ini memungkinkan tugas analitik data otomatis di atas database DICOMmetadata. Penulis mengklaim bahwa meskipun ada beberapa upaya sebelumnya dalam literatur, yaitu [8, 9], tidak ada yang memiliki tujuan untuk mengaktifkan kemampuan data mining (DM) yang sepenuhnya sewenang-wenang. Solusinya didukung oleh database relasional, yang model datanya perlu diperluas untuk mendukung Elemen Data baru. Itu juga menargetkan perbedaan antara atribut data's dari vendor yang berbeda dengan membuat pemetaan statis untuk atribut berbeda yang mewakili pengukuran yang sama. Namun, kerangka kerja ini tidak menyediakan antarmuka grafis untuk membuat laporan, peringatan, atau menjelajahi data. Dalam artikel yang diperbarui [21], penulis menunjukkan kesulitan pengindeksan sumber data yang sangat heterogen, seperti DICOM, yang mengakibatkan beberapa atribut DICOM tidak terindeks.

Mereka menyatakan bahwa Tidak hanya pendekatan SQL (NoSQL) mungkin diperlukan untuk menanganinya.

Arsitektur

Arsitektur yang diusulkan mengikuti model client-server klasik, tersegmentasi menjadi tiga lapisan yang berbeda: presentasi, bisnis, dan ketekunan. Dalam pola arsitektur ini, klien hanya bertindak sebagai lapisan presentasi dari sistem yang dikembangkan, menampilkan antarmuka kepada pengguna dan beralih ke server bila diperlukan. Pada gilirannya, server bertanggung jawab untuk mengimplementasikan lapisan bisnis dan persistensi aplikasi. Lapisan bisnis mencakup sebagian besar aplikasi's logika, menangani klien's permintaan dan memberikan respon yang memadai. Lapisan persistensi bertanggung jawab untuk menyimpan dan memelihara data di beberapa sesi.

Server'Fungsionalitas BI disediakan oleh tumpukan Python Scipy (yang mencakup pustaka NumPy, SciPy, dan pandas) dan pustaka scikit-learn. Metode yang dikembangkan menggunakan pustaka sebelumnya diekspos melalui antarmuka aplikasi (API) Representational State Transfer (REST), menggunakan Django dan toolkit kerangka kerja Django REST.[1] Lapisan persistensi didasarkan pada Sistem Manajemen Basis Data Relasional PostgreSQL (RDBMS), yang mempertahankan aplikasi's data yang berhubungan dengan logika, dalam hubungannya dengan panda's Hierarchical Data Format Store (HDFS), bertanggung jawab untuk penyimpanan data massal.

Klien adalah aplikasi Web yang mengikuti pola aplikasi halaman tunggal, di mana semua kode klien yang diperlukan (HyperText Markup Language (HTML), JavaScript (JS), dan Cascading Style Sheets (CSS)) dimuat dalam satu halaman dimuat. Aplikasi dikembangkan menggunakan kerangka kerja React, dengan bantuan wadah status Redux. Selain itu, menggunakan kerangka Bootstrap untuk mengelola sebagian besar antarmuka serta perpustakaan plotly.js untuk menangani rendering grafik.

Kerangka kerja ini memanfaatkan Dicoogle's fitur data mining untuk mengekstrak data dari objek DICOM di PACS. Komunikasi ini didukung oleh Dicoogle'layanan penemuan dan pengambilan konten, yaitu, titik akhir kueri layanan Web-nya. Pada gilirannya, layanan ini bergantung pada teknologi yang diterapkan oleh plug-in, seperti yang dijelaskan dalam [22].

Terlepas dari pengembangan dasbor, arsitektur yang disajikan, melalui lapisan bisnis dan persistensi, memungkinkan penyediaan data ke kerangka kerja pihak ketiga (misalnya, Kibana[2] atau Grafana[3]).

1 www.djangoproject.com

2 www.elastic.co/products/kibana

3 www.grafana.com

Angka 4 menyajikan modul fungsional yang akan dijelaskan di bagian selanjutnya, serta interaksinya.

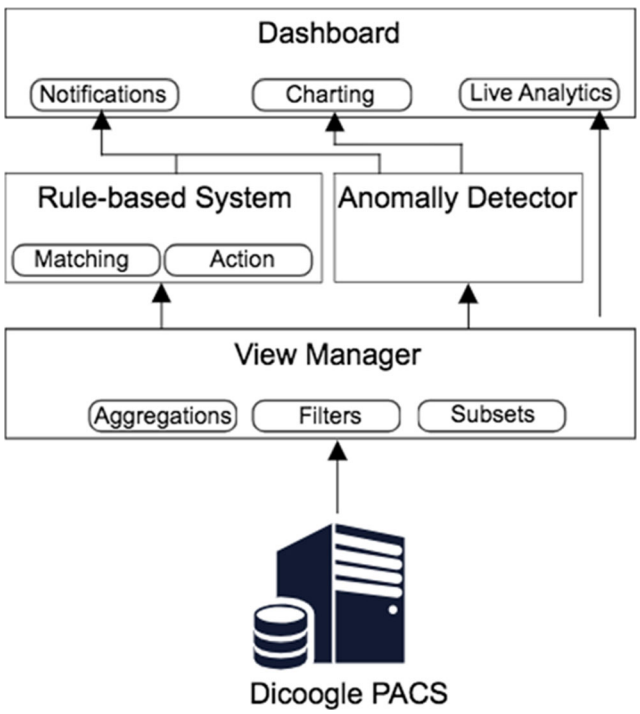
Pembersihan Data Berbasis Aturan

Salah satu perhatian utama adalah untuk memungkinkan manipulasi beberapa catatan tidak teratur secara bersamaan karena ketidakakuratan terjadi beberapa kali dalam dataset yang sama. Akibatnya, koreksi manual dari catatan tersebut akan menjadi tidak praktis. Untuk menghindari itu, sistem kontrol berbasis Aturan dirancang untuk mendukung fitur Pembersihan Data. Ia bekerja dalam dua fase: Pencocokan dan Tindakan. Pada fase pertama, aturan yang berbeda diterapkan dalam kumpulan data untuk mendeteksi catatan yang tidak akurat. Fase tindakan melakukan koreksi terhadap catatan.

Serangkaian aturan dan tindakan dasar dikembangkan untuk menyediakan layanan yang kuat dan menjaga, pada saat yang sama, kegunaan yang tinggi. Aturan mencakup sebagian besar kasus penggunaan yang ditentukan dalam konteks pekerjaan ini tetapi dapat diperluas untuk mencakup yang lain dengan menambahkan modul tambahan, yang ditulis dengan Python, yang ditafsirkan saat runtime.

Lima aturan berikut saat ini tersedia dalam sistem:

& Bidang Kosong: memungkinkan deteksi nilai kosong pada a kumpulan bidang. Masalah ini adalah anomali paling umum saat bekerja dengan data. Sayangnya, sebagian besar bidang ini tidak dapat disimpulkan, biasanya mengarah pada penghapusannya dari kumpulan data;



Gambar 4 Arsitektur kerangka kerja Intelijen Bisnis

- Nilai yang Diharapkan: untuk mendeteksi satu (atau lebih) nilai pada grup bidang tertentu. Ia bekerja dengan mendefinisikan satu set nilai dan bidang di mana nilai-nilai itu diharapkan;
- Ekspresi Reguler: versi yang lebih umum dan lanjutan dari aturan Nilai yang Diharapkan, yang membutuhkan Ekspresi Reguler' pengetahuan;
- Filter Date: memilih satu set rekaman yang bidang Tanggalnya cocok dengan interval yang ditentukan;
- Ekspresi: memungkinkan eksekusi pernyataan atau bahkan skrip kecil, memungkinkan untuk mendeteksi kasus yang lebih kompleks yang tidak mungkin dilakukan dengan aturan sebelumnya. Seharusnya hanya digunakan sebagai cadangan.

Aturan sebelumnya dapat digabungkan dengan tindakan berikut:

- Isi Bidang Kosong: menggantikan nilai kosong di bidang yang ditentukan dengan nilai yang dipilih;
- Ganti dengan Nilai: menggantikan serangkaian nilai yang ditentukan dalam rentang bidang tertentu;
- Ganti dengan Regex: kasus spesifik dari tindakan sebelumnya ketika nilainya, yang akan diganti, adalah
- Ekspresi Reguler; Date to Age: ini mengubah bidang tertentu (biasanya bidang Tanggal Lahir Pasien) ke Usia yang sesuai (paling sering bidang Usia Pasien) menurut standar DICOM;
- Normalisasi Usia: tindakan praktis untuk menormalkan bidang Usia, memberlakukan ukuran dalam Tahun, Bulan, Minggu, atau Hari;
- Ekspresi: ia menerima skrip python untuk mengimplementasikan tindakan gratis.

Penting untuk dicatat bahwa aturan yang ditetapkan mungkin tidak saling eksklusif karena aturan since's tindakan yang dipicu mungkin memperbarui nilai yang cocok dengan aturan lain. Untuk alasan ini, setiap aturan memiliki atribut prioritas, nilai integer, dan aturan dijalankan sesuai dengan prioritasnya.

Detektor Anomali

Komponen ini memungkinkan untuk mendeteksi anomali dalam data dengan mengaitkan bidang tertentu ke salah satu dari dua kategori: data ordinal atau nominal, berdasarkan bidang's tipe data. Seperti namanya, jika bidang yang dipilih'tipe data s adalah numerik (baik integer atau float), kemudian dilengkapi dengan deskripsi ordinal; jika tidak, itu diberikan dengan deskripsi nominal.

Deskripsi ordinal mengembalikan atribut hitungan, dengan jumlah total entri yang tidak kosong di bidang, serta beberapa statistik paling umum untuk serangkaian nilai numerik, seperti mean, standar deviasi, nilai minimum dan maksimum, dan persentil (25%, 50%, dan 75%). Di sisi lain, deskripsi nominal mengembalikan jumlah entri di mana nilai tertentu muncul. Juga, kedua deskripsi menampilkan nilai kosong' menghitung, jika diterapkan.

Sayangnya, deteksi otomatis bidang'Deskripsi s berdasarkan tipe datanya terkadang cacat. Seringkali, bidang terdiri dari kumpulan nilai bilangan bulat. Namun, nilai-nilai tersebut tidak mewakili variabel kontinu, yang berarti bahwa mereka seharusnya ditafsirkan sebagai kategori. Oleh karena itu, dimungkinkan untuk menerapkan deskripsi nominal secara manual untuk bidang yang seluruhnya terdiri dari nilai integer.

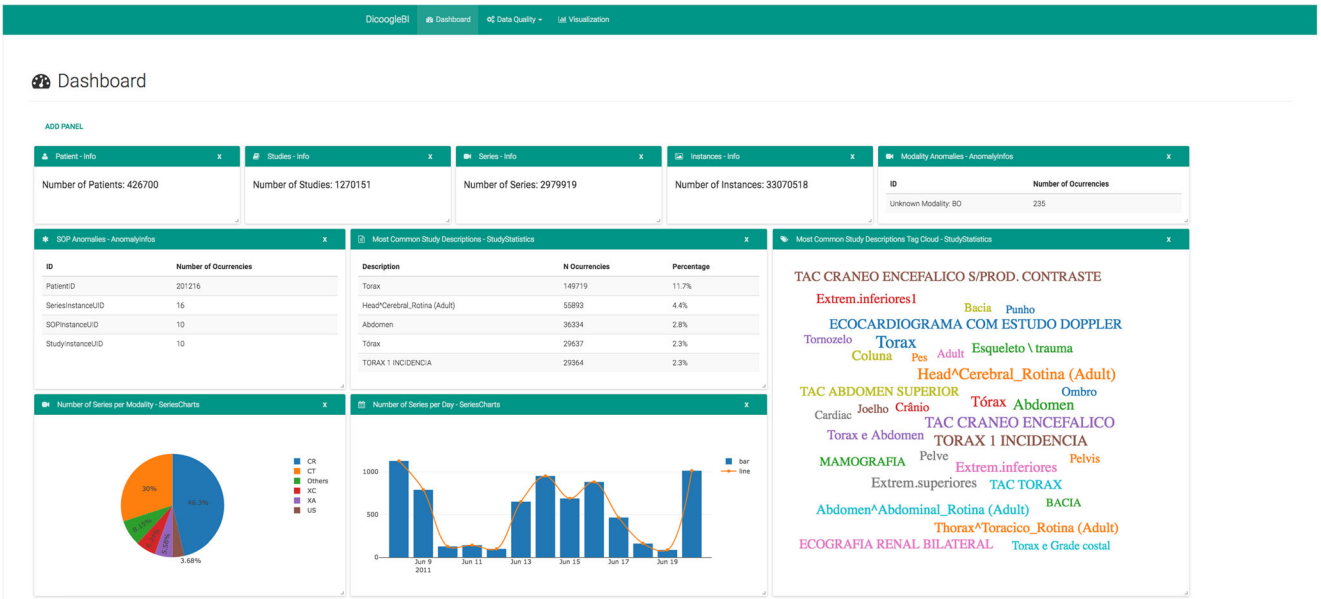
Lihat Manajer

Konsep Views diimplementasikan untuk memudahkan proses bekerja dengan kumpulan data yang besar dan heterogen. Sebagian besar kasus penggunaan tidak bekerja dengan semua data repositori, terutama, ketika ukuran dan heterogenitas data akan mempengaruhi kinerja sebagian besar operasi secara signifikan, bahkan yang paling dasar sekalipun. Selain itu, sangat sering analisis menargetkan subset tertentu dari repositori, seperti menganalisis data yang terkait dengan modalitas tertentu atau menganalisis laporan yang dilakukan dalam rentang tanggal tertentu. Mempertimbangkan hal ini, dikembangkan konsep tampilan yang bermaksud untuk mewakili kumpulan data yang lebih kecil dan lebih spesifik dari repositori PACS asli.

Oleh karena itu, Tampilan sangat penting karena menentukan konteks eksekusi komponen lain, yaitu, aturan dan pendeteksi anomali hanya akan diterapkan pada Tampilan yang ditetapkan, dan akibatnya kumpulan data. Ini memungkinkan Tampilan yang dibuat pengguna untuk mewarisi transformasi tampilan induknya, seperti konsep pewarisan Berorientasi Objek (OO). Tampilan yang dibuat pengguna dapat mewarisi baik dari Tampilan default (root) atau dari Tampilan buatan pengguna lainnya. Ini memberikan fleksibilitas dan ekstensibilitas ke sistem. Tampilan mendukung semua manipulasi data yang dirujuk sebelumnya. Untuk mendukung ini, sistem memungkinkan penggabungan tampilan'transformasi s dengan induknya's, sehingga mengoptimalkan penggunaan Dicoogle's Kemampuan Data Mining.

Untuk membuat tampilan baru, tiga transformasi disediakan:

- Agregasi: meniru grup SQL dengan operasi, mendefinisikan kumpulan bidang agregasi dan fungsi yang akan dilakukan. Dimungkinkan untuk menyediakan hanya satu fungsi sebagai parameter. Dalam hal ini, ini akan diterapkan ke semua bidang yang bukan bagian dari bidang agregasi. Namun, ini jarang menghasilkan output yang diinginkan. Untuk alasan ini, dimungkinkan juga untuk mendefinisikan satu atau lebih fungsi per kolom. Kemungkinan ini sangat berguna untuk mempertimbangkan Hirarki Entitas Informasi (Pasien, Studi, Seri, Gambar). Fungsi yang diperbolehkan adalah penjumlahan kumulatif, perkalian kumulatif, maksimum, minimum, median, simpangan baku, rata-rata, ukuran dan jumlah;
- Subset: memungkinkan pembuatan subset statis, mendefinisikan interval baik baris dan/atau bidang (yaitu, seleksi dan pembatasan proyeksi);
- Filter: memungkinkan untuk menggunakan aturan di atas, sebagai transformasi.



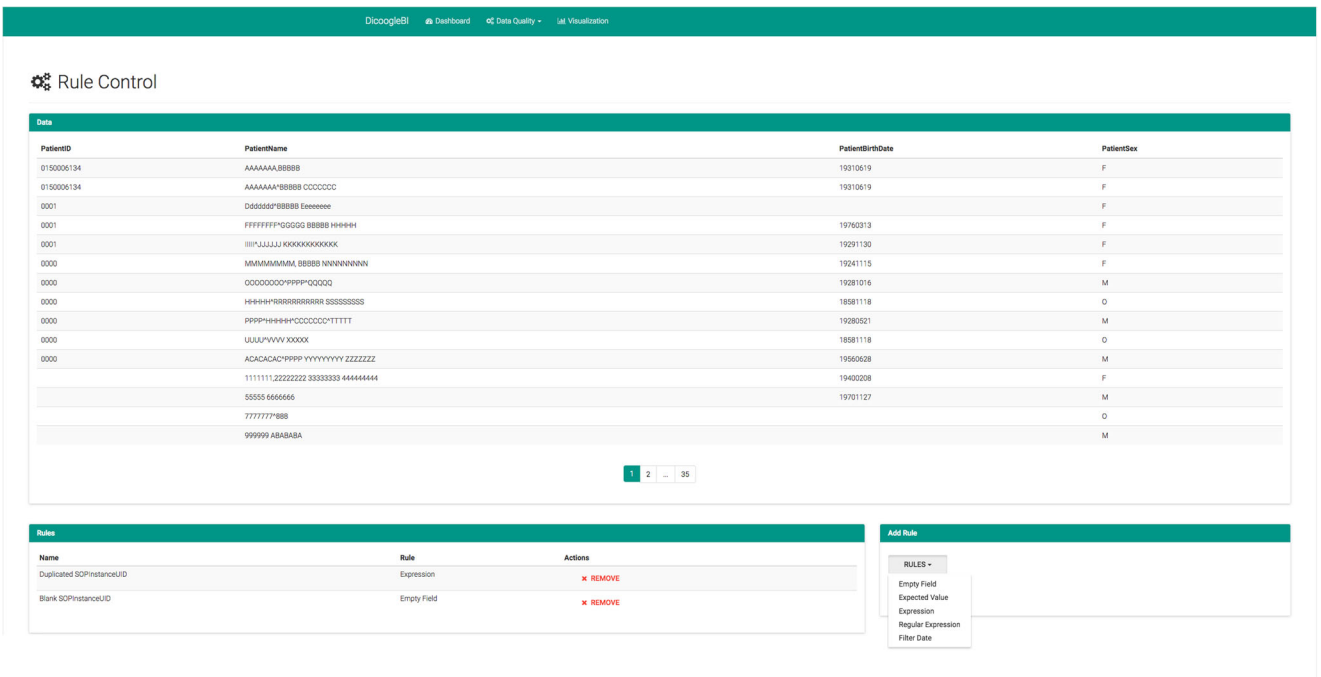
Gambar 5 Ikhtisar dasbor Dicoogle BI

Transformasi yang ditentukan dapat diterapkan menggunakan kemampuan penambangan data Dicoogle dalam dua situasi. Pertama, ketika filter Nilai yang Diharapkan ditentukan, itu diterjemahkan persis ke Dicoogle's bahasa permintaan. Misalnya, filter oleh Modality dan PatientSex, dan nilai yang diharapkan CR dan F masing-masing, akan diterjemahkan keBModalitas:CR DAN PatientSex:F`di Dicoogle'layanan permintaan.

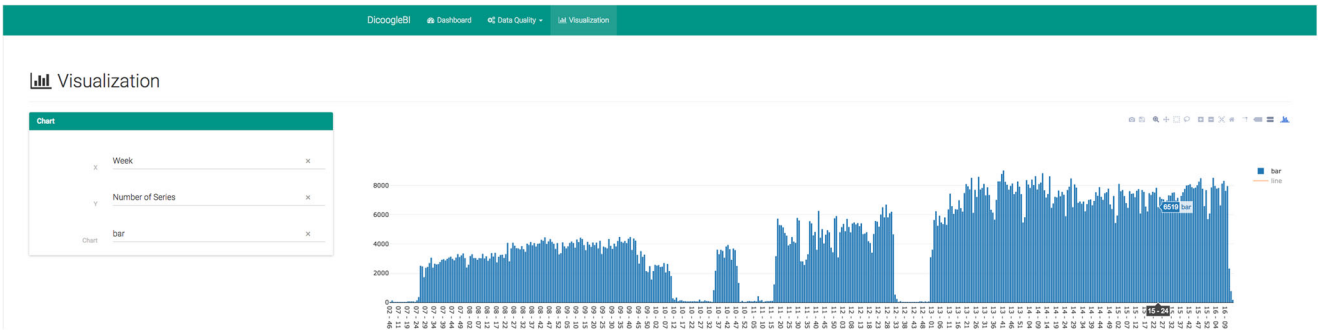
Kedua, ketika bidang transformasi subset didefinisikan, mereka juga dipetakan ke Dicoogle'antarmuka kueri. Misalnya, subset dengan bidang Modalitas dan Usia pasien, apakah t dijalankan ke kueri?'s kembali bidang = [Modalitas,Umur Pasien].

Dasbor

Solusi yang dikembangkan menyediakan kemampuan dasbor yang luas. Dasbor memungkinkan pengembangan halaman klien yang sepenuhnya dapat disesuaikan yang dapat mencakup salah satu komponen visualisasi yang tersedia, masing-masing dimasukkan ke dalam panel yang dapat diubah ukurannya dan diurutkan sepenuhnya. Visualisasi juga bergantung pada



Gambar 6 Contoh aturan yang diterapkan pada tampilan kumpulan data



Gambar 7 Ikhtisar antarmuka visualisasi Dicoogle BI

pandangan yang ditentukan. Jadi, sistem tidak hanya menyimpan tata letak dasbor, termasuk setiap panel's koordinat dan ukuran, tetapi juga instruksi yang diperlukan untuk mengisi komponen dengan data yang diinginkan.

Mengingat integrasinya dengan Dicoogle PACS yang mendasarinya, sistem yang diusulkan menyediakan kemampuan analitik waktu nyata. Ini berarti bahwa platform segera diberitahu ketika data baru tiba di arsip, menghindari keharusan membuat snapshot repositori, yang diperoleh secara berkala. Gambar baru secara otomatis ditambahkan ke Tampilan yang diperlukan dan dianalisis dengan benar oleh aturan yang ditentukan sebelumnya. Nantinya, notifikasi juga dikirim ke Dashboard dengan informasi terbaru.

Hasil

Bagian ini menunjukkan kemampuan penemuan konten dari kerangka kerja yang dikembangkan untuk mengekstraksi pengetahuan dari repositori pencitraan medis. Demonstrasi ini didasarkan pada arsip PACS lembaga afiliasi dengan sekitar 35 juta file. Dicoogle BI digunakan dan digunakan secara paralel dengan PACS institusional. Dalam lingkup pekerjaan ini, data sensitif dianonimkan menggunakan alat internal sehingga tidak akan diungkapkan.

Ikhtisar dasbor Dicoogle BI ditunjukkan pada Gambar. 5, termasuk beberapa widget. Di atas, ada empat widget informatif yang menghitung jumlah pasien (426700), studi (1270151), seri (2979919), dan contoh (33070518) dalam kumpulan data. Awalnya, kami melihat perbedaan antara nilai-nilai ini dan nilai-nilai sebelumnya yang dihasilkan oleh alat yang lebih awal dan lebih sederhana. Alat ini hanya melakukan penghitungan atribut pengenal gambar (35476975) dan pasien (279392).

Perbedaan antara jumlah gambar dan jumlah instans (pengidentifikasi) disebabkan oleh file DICOM yang digandakan dalam repositori. Hal ini dapat dijelaskan oleh fakta bahwa beberapa platform perangkat lunak membuat salinan file asli (misalnya, menghasilkan gambar dengan resolusi lebih rendah untuk tujuan tampilan cepat) tanpa memiliki SOPInstanceUID. diperbarui, yang melanggar standar, dan menyebabkan perkiraan yang terlalu tinggi sebelumnya dari jumlah gambar dalam repositori. Hal ini juga berdampak pada perkiraan jumlah pemangku kepentingan lainnya.

Dasbor juga menyertakan dua widget yang terkait dengan anomali dalam kumpulan data. Yang pertama menemukan Modalitas yang tidak diketahui, yaitu modalitas yang tidak didefinisikan dalam standar. Misalnya, diidentifikasi 235 seri berbeda dari modalitas yang disebut PL (Lainnya). Widget kedua mengidentifikasi pengidentifikasi duplikat dalam kumpulan data. Di tingkat pasien, juga diidentifikasi perbedaan yang disebabkan oleh pengidentifikasi ganda untuk catatan yang jelas berbeda. Misalnya, PatientID yang sama diberikan untuk gambar dengan tupel (PatientName, PatientBirthDate, PatientSex) yang berbeda. Dalam kasus ini, lebih dari 201.216 kasus yang teridentifikasi,[4] mewakili 0,6% dari instance repositori. Meskipun banyak PACS berpemilik dapat mengatasi ketidakkonsistenan ini, mereka tidak mendukung upaya analitis apa pun yang bergantung pada kualitas data yang dikumpulkan.

Widget lain dengan statistik juga disajikan di Dasbor. Misalnya, ada diagram lingkaran yang merangkum isi dataset menurut modalitas seri. Ada juga diagram batang dengan produksi seri per hari. Perhatikan penurunan produktivitas pada akhir pekan (11 dan 12 Juli, serta 18 dan 19 Juli) dibandingkan dengan hari kerja biasa. Terlihat juga dampak libur pada tingkat produktivitas institusi, dan penurunan yang tidak dapat dijelaskan pada hari Jumat dibandingkan dengan hari kerja lainnya.

Terakhir, widget tag cloud dengan deskripsi studi disertakan. Hal ini dimungkinkan untuk mengamati bahwa atribut StudyDescription memiliki kekurangan normalisasi. Statistik sebelumnya membutuhkan agregasi kumpulan data oleh SeriesInstanceUID dan kemudian oleh Modalitas sebelum fungsi hitungan dapat diterapkan.

Angka 6 menunjukkan antarmuka untuk mengonfigurasi aturan tampilan yang diberikan. Data sampel menunjukkan duplikat catatan PatientID yang dirujuk dalam paragraf sebelumnya. Catatan dikaburkan oleh alasan privasi, tetapi nama asli

4 PatientID yang sama diberikan untuk catatan dengan tupel PatientName, PatientBirthDate, PatientSex yang berbeda

akan selalu diganti dengan versi obfuscate yang sama. Dengan demikian, Anda dapat melihat PatientNames yang berbeda dengan pengenal yang sama. Tanggal Lahir Pasien yang sangat mencurigakan juga dapat dilihat, misalnya, 18581118.

Akhirnya, ikhtisar antarmuka pembuatan bagan aplikasi ditunjukkan pada Gambar. 7. Atribut dan tipe bagan dapat disesuaikan untuk tampilan tertentu. Bagan itu sendiri bersifat interaktif, memungkinkan untuk mengubah ukuran jendela visualisasi dan menerapkan transformasi ke sumbu. Fungsi grafis yang disajikan memungkinkan setiap pengguna untuk melakukan tugas DA dan BI pada konten PACS mereka bahkan jika mereka tidak memiliki keterampilan pemrograman.

Kemampuan Dicoogle'Modul BI s juga dapat diterapkan ke PACS pihak ketiga lainnya berkat Dicoogle's kemampuan mengindeks konten arsip PACS lainnya. Selain itu, dapat digunakan untuk mendukung upaya penelitian ini di PACS nonproduksi. Meskipun, manfaat memiliki saluran analitis waktu nyata akan hilang dalam kedua kasus ini.

Kesimpulan

Saat ini, eksploitasi laboratorium pencitraan medis telah menjadi bagian penting dari model bisnis institusi kesehatan. Repositori pencitraan medis produksi menyimpan sejumlah besar data. Karena data ini berasal langsung dari praktik medis, data ini memiliki akurasi ekstrem untuk tujuan analisis. Eksploitasi tepat waktu dari data ini telah menjanjikan untuk meningkatkan efisiensi praktik bisnis institusi medis, serta kualitas layanan kesehatan. Meskipun pentingnya ini telah diakui oleh masyarakat, volume dan tingkat produksi data pencitraan medis membuat analisis manual menjadi tidak praktis.

Kerangka kerja Intelijen Bisnis Dicoogle, yang dijelaskan dalam dokumen ini, mengatasi masalah ini. Ini memungkinkan pengembangan analisis otomatis' alur kerja dilakukan langsung di atas repositori institusional langsung tanpa meminta pembuatan gudang data khusus. Manajer bisnis dan peneliti layanan kesehatan dapat secara otomatis memperoleh pengetahuan melalui repositori besar, yang sebelumnya akan memakan terlalu banyak waktu. Selain itu, ia memiliki kemampuan untuk meningkatkan kualitas repositori dengan menggunakan sistem aturan lengkapnya yang menyediakan semua fungsi Pembersihan Data yang diperlukan, serta kontrol Tampilan yang serbaguna. Alat-alat ini dapat dilengkapi dengan komponen analisis data yang disediakan, seperti grafik dan modul deteksi anomali. Akhirnya, komponen-komponen ini dapat digabungkan lebih lanjut pada aplikasi's Dasbor, memungkinkan operator untuk lebih menyesuaikan alur kerjanya sendiri.

Informasi Pendanaan Tiago Marques Godinho didanai oleh Fundação para a Ciência e Tecnologia (FCT) berdasarkan perjanjian hibah SFRH/BD/104647/2014. Rui Lebre menerima dukungan dari Program Terpadu SR&TDBSOCA^ (Ref.CENTRO-01-0145-FEDER-000010), didanai bersama oleh program Centro 2020, Portugal 2020, Uni Eropa, melalui Dana Pembangunan Regional Eropa. Pekerjaan ini memiliki menerima dukungan dari ERDF European Regional Development Fund melalui Program Operasional untuk Daya Saing dan Internasionalisasi, Program COMPETE 2020, dan oleh Dana Nasional melalui FCT, Fundação para a Ciência ea Tecnologia dalam proyek PTDC/EEI-ESS/6815/2014.

Referensi

1. Hamilton B: Data besar adalah masa depan perawatan kesehatan. Teaneck: Sadar, 2012
2. Godinho TM, Viana-Ferreira C, Bastiao Silva LA, Costa C: Mekanisme perutean untuk outsourcing cloud repositori pencitraan medis. Informasi Kesehatan IEEE J Biomed 20(1):367-375, 2016. https://doi.org/10.1109/JBHI.2014.2361633
3. Santos M, Bastiao L, Costa C, Silva A, Rocha N: BPenambangan Data Klinis di Rumah Sakit Kecil PACS: Kontribusi untuk Peningkatan Departemen Radiologi,^ dalam Sistem dan Teknologi Informasi untuk Meningkatkan Perawatan Kesehatan dan Sosial. Hershey: IGI Global, 2013, hlm. 236-251
4. Mildenberger P, Eichelberg M, Martin E. Pengantar standar DICOM. Eur Radiol 12(4):920-927, 1, 2002. https://doi. org/10.1007/s003300101100
5. Digital Imaging and Communications in Medicine (DICOM) Bagian 3: Definisi objek informasi, NEMA, Standard, 2017.
6. Digital Imaging and Communications in Medicine (DICOM) Bagian 7: Pertukaran Pesan, NEMA, Standard, 2017.
7. Pianykh OS: Digital Imaging and Communications in Medicine (DICOM): pengantar praktis dan panduan bertahan hidup, Vol. 26. Berlin: Springer Science & Business Media, 2009, 424 hal
8. Wang S, Pavlicek W, Roberts CC, Langer SG, Zhang M, Hu M, Morin RL, Schueler BA, Wellnitz CV, Wu T: Database DICOM otomatis yang mampu melakukan penambangan data sewenang-wenang (termasuk indikator dosis radiasi) untuk pemantauan kualitas. Pencitraan Digit J 24 (2): 223-233, 2011. https://doi.org/10.1007/s10278-010-9329-y
9. HuM, PavlicekW, Liu PT, ZhangM, Langer SG, Wang S, Place V, Miranda R, Wu TT: Informatika dalam radiologi: metrik efisiensi untuk produktivitas perangkat pencitraan. RadioGraphics 31(2):603-616, 2011. https://doi.org/10.1148/rg.312105714
10. Ondategui-Parra S, Erturk SM, Ros PR: Survei penggunaan indikator kualitas di departemen radiologi akademik. Am J Roentgenol 187(5):W451-W455, 1, 2006. https://doi.org/10.2214/AJR.05. 1064
11. Raghupathi W, Raghupathi V: Analisis data besar dalam perawatan kesehatan: janji dan potensi. Sistem Informasi Ilmu Kesehatan 2(1):3, 1, 2014. https://doi.org/10.1186/2047-2501-2-3
12. Nagy PG, Warnock MJ, Daly M, Toland C, Meenan CD dkk.: Informatika dalam radiologi: dasbor grafis berbasis web otomatis untuk intelijen bisnis operasional radiologi. RadioGraphics 29(7):1897-1906, 1, 2009. https://doi.org/10.1148/ rg.297095701
13. Andreu-Perez J, Poon CCY, Merrifield RD, Wong STC, Yang GZ: Data besar untuk kesehatan. Informasi Kesehatan IEEE J Biomed 19(4):1193-1208, 2015. https://doi.org/10.1109/JBHI.2015.2450362
14. Viceconti M, Hunter P, Hose R: Data besar, pengetahuan besar: data besar untuk perawatan kesehatan yang dipersonalisasi. Informasi Kesehatan IEEE J Biomed 19(4): 1209-1215, 2015. https://doi.org/10.109/JBHI.2015.2406883
15. Langer SG: Tantangan untuk penyimpanan data dalam penelitian pencitraan medis. J Digit Pencitraan 24(2):203-207, 2011. https://doi.org/10. 1007/s10278-010-9311-8

16. Watson HJ, Wixom BH: Keadaan intelijen bisnis saat ini. Komputer 40(9):96-99, 2007. https://doi.org/10.1109/MC.2007.331
17. Chen H, Chiang RHL, Storey VC: Intelijen dan analitik bisnis: dari data besar hingga dampak besar. MIS Q 36(4):1165-1188, 2012
18. Langer SG: Arsitektur database yang fleksibel untuk menambang objek DICOM: gudang data DICOM. Pencitraan Digit J 25(2):206-212, 2012. https://doi.org/10.1007/s10278-011-9434-6
19. JTL Wang, MJ Zaki, HTT Toivonen, dan D. Shasha, Pengantar Data Mining di Bioinformatika. Dalam: Data Mining in Bioinformatics, Springer, London, 2005, hlm. 3-8. https://doi.org/10.1007/1-84628-059-11.
20. Valente F, Silva LAB, Godinho TM, Costa C: Anatomi PACS open source yang dapat diperluas. Pencitraan Digit J, 2015.https://doi.org/10.1007/s10278-015-9834-0
21. Costa C, Freitas F, Pereira M, Silva A, Oliveira JL: Mengindeks dan mengambil data DICOM dalam arsip yang tersebar dan tidak terstruktur. Int J Comput Assist Radiol Surg 4(1):71-77, 1, 2009. https://doi.org/10. 1007/s11548-008-0269-7
22. Bastiao L, Santos M, Costa C, Silva A, Rocha N: Statistik Dicoogle: Menganalisis efisiensi dan kualitas layanan laboratorium pencitraan digital. Heildelberg: Springer, 2013
23. Santos M, Bastiao L, Costa C, Silva A, Rocha N: Dicom dan penambangan data klinis di pacs rumah sakit kecil: Sebuah studi percontohan. Dalam: Konferensi Internasional tentang Sistem Informasi ENTERprise. Berlin: Springer, 2011, hlm. 254-263