

Исследование способов хранения, передачи и обработки больших объёмов данных в DICOM формате для медико-информационных систем

Alexandr GOLUBEV, Peter BOGATENCOV, Grigore SECRIERU
Research and Educational Networking Association of Moldova
galex@renam.md

Абстракт— В работе анализируются возможные решения проблем хранения и доступа к медицинским изображениям на основе распределённого хранилища больших объёмов данных с разным уровнем доступа. Уровень доступа в данном случае характеризуется как уровнем безопасности, так и скоростью получения информации для различных типов исследований.

Современные медицинские информационные системы не могут обойтись без интеграции с медицинским оборудованием. Статья посвящена рассмотрению проблем и решения для оптимизации получения, обработки и хранения данных медицинских исследований в области радиологии. Стандартом в медицинских изображениях является формат DICOM, позволяющий сохранять исследования в хорошем качестве совместно с персональными данными пациента. Проблемы возникают из-за резкого возрастания объёма информации так как одно исследование может занимать до 1 гигабайта и состоять из тысяч изображений.

В 2015 году в Молдове был запущен проект «DICOM Network» целью которого было обеспечение доступа к исследованиям для каждого врача или пациента. В результате проекта к системе было подключено 11 медицинских аппаратов, производящих обследования в формате DICOM. Система собирает и обрабатывает более 500 гигабайт данных в месяц. В данной статье произведён анализ возможных решений для рационального архивирования таких больших объёмов данных.

Ключевые слова — DICOM; VI-SEEM; RENAM; Distributed System; HPC; Processing Algorithms; Radiology; e-Health; Medical Emergency

I. ВВЕДЕНИЕ

Современные медицинские информационные системы объединяют различные виды медицинского оборудования. В статье описываются актуальные проблемы и способы их решения для оптимизации обработки и хранения медицинских рентгенологических исследований. Стандартом для работы с медицинскими изображениями является формат DICOM, который позволяет хранить исследования в хорошем качестве с включёнными персональными данными пациента. Основная проблема хранения данных в формате DICOM [1] вызвана тем, что одно исследование может занимать более 1 гигабайта и состоит из тысяч изображений.

Проблема хранения архивов медицинских исследований на национальном уровне может рассматриваться как работа с Big Data. Решение этой проблемы должно учитывать различные уровни доступа к данным. С одной стороны, медицинское расследование содержит личные данные пациента, что означает, что доступ к данным должен быть ограничен и защищён. Это может быть достигнуто на основе разрешённых категорий пользователей для доступа к индивидуальным обследованиям. С другой стороны,

данные должны быть доступны многим пользователям, например, пациенту или врачу, из любого местоположения в режиме on-line. Одним из главных приоритетов является производительность системы, которая должна обеспечивать высокоскоростной доступ к огромному объёму архивированных данных.

Таким образом процесс сохранения и передачи больших объёмов медицинских данных можно разделить на несколько составляющих: архивирование и хранение, поиск и доступность данных, передача данных на компьютер пользователя и визуализация данных в клиентском приложении.

После осуществления обследования и импорта набора DICOM изображений из оборудования возникает проблема архивирования и соответственно хранения данных [2,3]. Если снимок рентгена обычно не содержит более 2-3 изображений, то топографическое обследование может содержать до 1000 срезов и занимать до нескольких гигабайтов памяти на физическом диске. Несложно посчитать что для большой больницы или крупного диагностического центра с ежедневным потоком в 500 и более пациентов объёмы данных будут исчисляться терабайтами в месяц, а архив обследований по закону должен храниться минимум 3 года. В результате медицинские

учреждения могут самостоятельно не справиться с таким объёмом данных, не иметь достаточно мощности для их хранения.

При хранении данных обязательно нужно учитывать безопасность доступа к данным и исключить утечку персональных данных пациентов, так как каждый файл содержит персональную информацию, включая имя фамилию и прочие личные данные. Это важно учитывать при распределённом хранении данных и особенно при хранении данных на внешних системах за пределами медицинского учреждения.

Стоит учесть, что при решении практической задачи импорта медицинских изображений в формате DICOM непосредственно с устройства, приходится создавать локальные сети между устройством и системой хранения данных, изолированные от глобальных сетей, так как подключение медицинского оборудования непосредственно к сети интернет небезопасно и не рекомендуется производителями. Таким образом на каждом этаже или даже в кабинете с оборудованием предполагается размещение некоего процессора для импорта данных для записи их в глобальную базу данных. В виду вышеизложенных факторов, хранение данных на одном элементе памяти не представляется возможным или будет слишком неэффективным.

Другой проблемой является передача данных по каналам Internet непосредственно конечному пользователю. Конечно при обследовании в 2-3 среза пакет данных не превышает 1-2 мегабайт, но при передаче полноценного топографического обследования в гигабайт и более скорость загрузки можем составлять десятки минут, в зависимости от скорости канала, что неприемлемо для конечных пользователей. Ещё более критичным является использование этих данных непосредственно в операционной, когда скорость является еще более важным компонентом. К проблеме передачи больших объёмов данных добавляется еще проблема локальной обработки данных при визуализации, особенно если визуализация проходит на мобильных устройствах с ограниченными вычислительными способностями.

В данной статье рассматриваются алгоритмы и способы решения проблемы работы с большими объемами данных, а также рассмотрены практические пути из решения на базе приложения "DICOM Network". Также рассматриваются перспективы международного сотрудничества для решения проблемы хранения и доступа к данным на базе платформы Европейского проекта VI-SEEM [4].

II. DICOM NETWORK

Проект «DICOM Network» был запущен в Молдове в 2015 году. Цель проекта - предоставить доступ к исследованиям для медицинского персонала с соответствующими правами доступа, а также пациентов к персональным исследованиям радиологии. В результате проекта к системе было подключено одиннадцать медицинских приборов. Сегодня система собирает и обрабатывает более 500 гигабайт данных в месяц.

Компоненты систем хранения данных и обработки данных распределены между различными процессорами и хранилищами, которые могут быть настроены с использованием определенных интерфейсов. Общая архитектура системы «DICOM Network» представлена на рисунке 1.

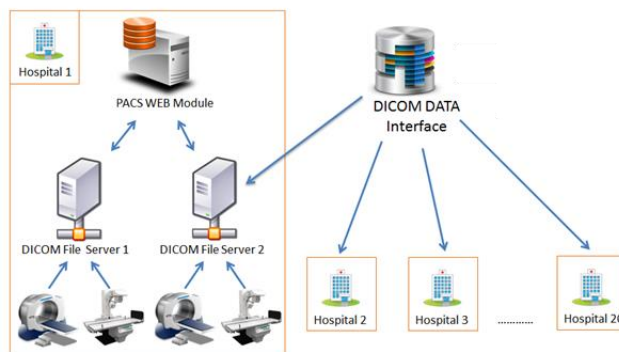


Рисунок 1. Общая схема сетевой архитектуры DICOM

Структура системы включает следующие серверы и модули данных:

- Данные из оборудования собираются платформой «DICOM Server», которая может быть установлена в том же учреждении с подключением используемого медицинского оборудования или эти сервера могут быть распределены между другими учреждениями, городами или странами.
- Все исследования (DICOM Images) архивируются на серверах DICOM, но информация о исследованиях хранится в базе данных DICOM Portal (например, www.dicom.md). Различные службы DICOM подключены к одному DICOM-порталу.
- DICOM Portal хранит все данные, такие как информация пользователя, информация о доступе, системные настройки, настройки сервера DICOM и некоторые другие, но не изображения DICOM, которые он использует. Каждое учреждение может самостоятельно развертывать DICOM Portal с использованием собственных ресурсов.
- Интерфейс DICOM DATA собирает информацию о пользователях и исследованиях из всех порталов DICOM и предоставляет функциональные возможности для обмена и унификации данных.

Система охватывает все необходимые рабочие процессы для обработки и документирования медицинских исследований - от сбора изображений обследования в медицинской карте пациента. «DICOM Network» предлагает расширенную функциональность для повышения качества медицинского управления и обеспечения доступа к исследованиям. Это помогает врачам, специалистам и другим авторизованным пользователям получить доступ к структурированной базе данных медицинских изображений, позволяет документировать изображения, которые собираются из

различных медицинских аппаратов. На институциональном уровне система помогает снизить затраты на обследование, повысить качество предоставляемых услуг.

DICOM Network уже запущен в производство и портал доступен по ссылке <http://www.dicom.md/>. Интерфейс представлен на рис. 2.



Рисунок 2. Портал Dicom Network

Первоначально система была развернута в Национальном центре скорой помощи Молдовы, и в течение первого года функционирования она показала привлекательность для персонала и эффективность, которая отображена в следующих цифрах:

- Запущено 2 портала DICOM: <http://dicom.md/> - национальный портал для молдавского сообщества и <http://viseem.dicom.md/> для международного исследовательского сообщества.
- Установлено 3 DICOM Servers
- 11 типов медицинского оборудования подключено к системе "DICOM Network".
- 16 специалистов по радиологии работают с системой в режиме 24/7.
- Около 250 обследований в день импортируются в систему.
- Более 500 врачей имеют доступ к системе непосредственно с их рабочего места.
- Более 15 000 Обследований в год собираются в системе.
- Более 135 000 Обследований собрано в системе.

"DICOM Network" активно развивается, так как врачи заинтересованы в получении доступа к радиографическим обследованиям непосредственно со своего рабочего места, а руководство в экономии средства на расходные материалы [5]. Подключается новое оборудование что увеличивает количество импортируемых обследований. Основной проблемой с которой сталкивается система – это постоянно растущие объемы данных которые необходимо сохранять.

На текущий момент данные сохраняются на 2-х

DICOM серверах и одном элементе памяти. Причём на DICOM серверах хранится архив за последний месяц, а на общем элементе памяти общий архив за последние 2 года работы системы. Данная архитектура позволяет обеспечить наиболее быстрый доступ к более свежим данным, которые используются чаще. Загрузка с общей памяти происходит значительно дольше.

III. ОПТИМИЗАЦИЯ ДАННЫХ В ФОРМАТЕ DICOM

Для описания решения по оптимизации DICOM данных для начала рассмотрим алгоритмы и процессы, которые необходимы для любой информационной системы работающий с радиографией. Для получения исходных снимков DICOM предусматривает импорт данных непосредственно от оборудования. После импорта изображения необходимо его обработать и сохранить в элементе памяти. В результате составляется база данных медицинских исследований доступная пользователю. Пользователь в свою очередь при обращении к БД должен получить данные по локальной или глобальной сети, после чего при помощи программы визуализации открыть DICOM изображения на своём рабочем месте.

Таким образом проблема оптимизации работы с данными в системе можно разделить на 3 этапа. Во-первых, при импорте и записи исходных файлов в память необходима архивация данных для максимального уменьшения объёма при полном сохранении качества изображений. Во-вторых, при обращении к данным необходимо выдать пользователю данные как можно быстрее и при этом оптимизировать формат для уменьшения объёма передаваемых изображений. В-третьих, данные должны быть оптимизированы для ускорения загрузки и обработки их в локальном приложении (процессоре).

IV. АРХИВИРОВАНИЕ ДАННЫХ

Как уже было сказано выше одно радиографическое обследование может варьировать от 1-го до 1000 изображений. При этом размер может достигать 1 и более гигабайт. Разумеется, наиболее востребованным является архивация больших томографических обследований с большим количеством срезов.

Для начала рассмотрим структуру файла DICOM. В общих чертах каждый файл состоит из meta data пациента и самого изображения в RAW формате. Так как при обработке вся мета информация записывается в базу данных, то при архивации мы можем отбросить мета данные и сохранять лишь само изображение. При необходимости возможно восстановить эти данные и сформировать DICOM файл при запросе клиента или же использовать новый формат файла при передаче специализированному приложению DICOM VIEWER. Конечно мета данные составляют лишь малую часть файла и занимают всего пару килобайт, но стоит учитывать, что эти данные для одного обследования практически идентичны, таким образом для набора данных в 1000 срезов можно будет экономить значительные объёмы физического пространства. С

другой стороны, исключение персональных данных пациента из исходных файлов значительно уменьшает возможность утечки данных, а также позволяет передавать не персонализированные данные.

Архивирование изображения намного более сложная задача, которая требует более сложных алгоритмов сжатия данных. Стоит учитывать, что необходимо исключить потерю качества изображения так как каждый пиксель и оттенок важны для обработки и визуализации данных. В ходе анализа было установлено, что архивация индивидуального среза без потери качества изображения не даёт ощутимого результата, но стоит отметить что доля индивидуальных срезов достаточно мала и относится в основном к рентгену, когда одно обследование состоит из 1-2 изображений. Доля объёма данных такого типа в памяти не превышает 1%. При анализе более сложных томографических обследований с 100 и более срезами было установлено что для их архивирования можно применить стандартные алгоритмы сжатия видеоизображений, так как изменения между срезами (кадрами) несущественны - на рис. 3 показаны «соседние» 9 срезов.

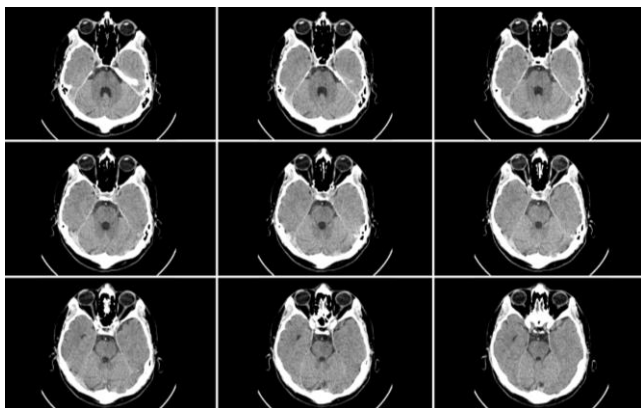


Рисунок 3. Аналогичные лучевые исследования

Как видно из рисунка 3 различия в соседних срезах столь малы, что можно с уверенностью преобразовать эти срезы в видеоряд где сохранять лишь различия для картинок, а не RAW формат для каждого изображения. Таким образом можно уменьшить объем изображений со множественными срезами в 10 и более раз. Конечно необходимо учитывать, что для архивации понадобятся значительные вычислительные мощности, что предполагает использование высокопроизводительных систем типа HPC или GRID.

V. ОПТИМИЗАЦИЯ ДОСТУПА К ДАННЫМ

Другим проблемным аспектом хранения данных является распределение данных по различным элементам памяти с разным уровнем безопасности и скоростью доступа. Как уже было сказано выше, хранение данных на одном элементе памяти невозможно по целому ряду причин. Для решения данной проблемы было предложена распределённая система хранения данных изображённая на рис. 4:

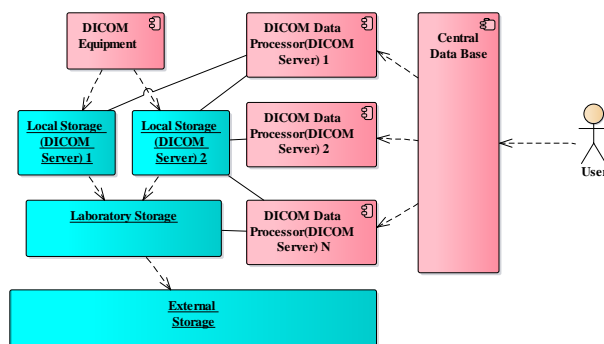


Рисунок 4. Распределённая система хранения данных.

Основными компонентами предлагаемой распределённой системы являются [6]:

- **Local Storage** – так как сервер подключений непосредственно к оборудованию должен находиться как в локальной сети так и в глобальной сети, а медицинское оборудование лишь в локальной сети, то для решения данной проблемы зачастую используют локальный сервер физически расположенный рядом с оборудованием, то есть один на отделение или этаж. Именно этот сервер принимает DICOM данные и выполняет первичную обработку данных. С одной стороны, данный сервер обеспечивает наиболее быстрый доступ к данным, но с другой объём памяти у данного сервера редко превышает 1-2 терабайт что предполагает хранение данных в течении 1-2 месяцев. Это позволяет обеспечить высокоскоростной доступ к последним данным. По истечению предопределённого периода данные архивируются и пересылаются на “Laboratory Storage” или “External Storage”.
- **Laboratory Storage** – элемент памяти для одного учреждения, доступный по локальной сети. Данный storage ограничен финансовыми возможностями учреждения и может хранить данные в течении 1-3 лет.
- **External Storage** – элемент памяти, расположенный за пределами организации где были выполнены исследования и доступный лишь по глобальной сети internet. Данный тип памяти конечно самый низкосортной, но зато может централизованно хранить огромный архив данных на национальном или региональном уровне.

VI. ОПТИМИЗАЦИЯ ПЕРЕДАЧИ ДАННЫХ.

Независимо от места хранения данных для их визуализации изображения должны быть загружены на компьютер конечного пользователя. При этом независимо от канала передачи, которым может быть, как локальная сеть с высокой пропускной способностью, так и глобальная сеть интернет с низкими скоростями, пакеты данных могут быть очень большими, а от скорости загрузки может зависеть качество медицинских услуг, удобство и заинтересованность врача в использовании информационной системы.

Основным решением безусловно является система

архивирования и разархивирования данных описанная выше, но стоит учитывать и следующие возможности для оптимизации передачи:

- Подготовка пакета данных с учётом разрешения принимающего клиентского приложения. То есть если пользователь просматривает обследование на мобильном устройстве с низким разрешением нет смысла передавать полноэкранные изображения в максимально разрешении.
- Возможность загрузки конкретного среза в максимально разрешении. По запросу клиента нужно предусмотреть возможность отсылки конкретного изображения в максимально разрешении, например, при необходимости zoom in/zoom out.
- Погрузка данных в background отдельным потоком. При этом конечный пользователь сможет начать визуализацию не полного набора данных, в то время как полный набор будет подгружаться асинхронно и отображаться по мере погрузки.
- Кэширование данных на уровне клиента и сервера.

Всё вышперечисленное может существенно ускорить процесс передачи данных и сделает приложение для визуализации более удобным.

VII. ПРЕДПОДГОТОВКА ДАННЫХ ДЛЯ ВИЗУАЛИЗАЦИИ

После загрузки данных на компьютер конечного пользователя приложение, осуществляющее визуализацию DICOM изображений, должно обработать данные для отображения. При этом все данные должны быть загружены в оперативную память для быстрой обработки. Стоит отметить что DICOM Viewer (приложение, отображающее DICOM изображения) должен не только показывать изображение и менять срезы, но также и осуществлять более сложные операции от рисования до построения 3D модели и моделирования тканей. Не каждый персональный компьютер, а тем более мобильное устройство, обладает удовлетворительными вычислительными возможностями, таким образом вышперечисленные операции могут занять много времени.

Решением данной проблемы должна стать пред подготовка данных для визуализации еще на сервере. При условии использования высокопроизводительных систем, сервер может построить и передать готовые модели, которые не будут требовать сложных преобразований в клиентском приложении.

Выводы

На текущий момент "Dicom Network" хоть и активно развивается, но не полностью реализован заложенный в систему потенциал. С учётом растущего количества медицинского оборудования, тенденции к модернизации и компьютеризации медицинских учреждений система сможет получать и должна будет

обрабатывать десятки терабайт исходной информации. Хранение и последующая передача таких больших объёмов данных является дорогостоящим процессом, невозможным без оптимизации. С другой стороны, для успешного развития системы необходимо предоставить архив не на 3 года, как это предусмотрено по закону, а на десятки лет для отслеживания состояния больного и ведения полноценной медицинской карты. Стоит учитывать и необходимость наличия резервных копий столь важной информации. Несложно подсчитать что даже для такой небольшой страны как Молдова объёмы данных слишком велики чтобы хранить их в необработанном виде. Таким образом вопрос оптимизации и архивирования данных является ключевым фактором для развития подобного рода систем.

Сбор данных в системе «DICOM Network» предоставит огромные возможности для исследователя. Хотя система сейчас находится на стадии эксплуатации, функциональность «DICOM Network» постоянно улучшается. В процессе реализации системы бенефициары определили свои потребности в предоставлении дополнительных функций и услуг, таких как:

- Изучение и реализация новых методов оптимизации передачи и архивирования данных.
- Предварительная обработка изображений и обнаружение аномалий.
- Включение экспертных систем диагностики в помощь врачу.
- Разработка открытых API для «Dicom Network» для сбора, архивирования, доступа и обработки медицинских изображений на международном уровне.

Алгоритмы и решения, рассмотренные в данной статье, могут быть применены в любой медицинской информационной системе и не привязаны к конкретному проекту, так как разработанный подход предполагает тесную интеграцию со стандартом DICOM и скорее дополняет его нежели изменяет.

Благодарности

Данная работа выполнена при поддержке проекта Европейской Комиссии VI-SEEM (VRE for regional Interdisciplinary communities in Southeast Europe and the Eastern Mediterranean, грант контракт 675121).

Литература

- [1] DICOM format description - <http://dicom.nema.org/standard.html>
- [2] Alexandr Golubev, Peter Bogatencov, Grigore Secrieru, Nicolai Iliuha. DICOM Network - Solution for Medical Imagistic Investigations Exchange. International Workshop on Intelligent Information Systems. Proceedings IIS. 13-14 September, Chisinau, IMI ASM, 2011, pp. 179-182. ISBN 978-9975-4237-0-0

- [3] A. Anagnostaki, S. Pavlopoulos, E. Kyriakou, and D. Koutsouris, "A Novel Codification Scheme Based on the VITAL and DICOM Standards for Telemedicine Applications," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 12, pp. 1399–1411, 2002.
- [4] H2020 VI-SEEM project - <https://vi-seem.eu/>
- [5] S. N. Becker, R.L. Arenson. Cost and benefits of Picture Archiving and Communication System. *Journal of the American Medical Informatics Association*, Volume 1, pp. 361-371, Number 5, Sep / Oct 1994
- [6] Peter Bogatencov, Nicolai Iliuha, Grigore Secrieru, Alexandr Golubev. DICOM Network for Medical Imagistic Investigations Storage, Access and Processing. "Networking in Education and Research", *Proceedings of the 11th RoEduNet IEEE International Conference*, Sinaia, Romania, January 17-19, 2013, pp. 38-42. ISSN-L 2068-1038