Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра математической теории
упругости и биомеханики

Анализ «больших данных» в сфере государственного управления.

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы

направления (специальности) 09.03.03 Прикладная информатика

механико-математический факультет

Щепеткова Максима Алексеевича

Научный руководитель
кандидат ю.н., доцент _____ Р.В. Амелин
подпись, дата

Зав. кафедрой
доктор ф.-м.н., профессор _____ Л.Ю. Коссович
подпись, дата

Саратов 2018

ВВЕДЕНИЕ

Цифровые технологии пронизали жизнь современного человека. Объем данных о самых разных сторонах жизни растет, и одновременно растут возможности хранения информации.

Большинство экспертов сходятся во мнении, что ускорение роста объема данных является объективной реальностью. Социальные сети, мобильные устройства, данные с измерительных устройств, бизнес-информация – вот лишь несколько видов источников, способных генерировать гигантские объемы информации.

Значительную часть информации создают не люди, а роботы, взаимодействующие как друг с другом, так и с другими сетями данных – такие, как, например, сенсоры и интеллектуальные устройства. При таких темпах роста количество данных в мире, по прогнозам исследователей, будет ежегодно удваиваться. Количество виртуальных и физических серверов в мире вырастет десятикратно за счет расширения и создания новых data-центров. В связи с этим растет потребность в эффективном использовании и монетизации этих данных. Поскольку использование Big Data в бизнесе требует немалых инвестиций, то надо ясно понимать ситуацию. А она, в сущности, проста: повысить эффективность бизнеса можно сокращая расходы или/и увеличивая объем продаж.

Актуальность моей работы заключается в том, что в современном мире развивающихся технологий хранения информации крайне важно использовать эти огромные пласты информации эффективно, то есть применять результаты анализа данных так, чтобы оптимизировать PR коммуникации и как следствие поспособствовать развитию того или иного предприятия или компании.

Цель работы: выявить особенности использования Big Data в здравоохранении.

Для реализации указанной цели в работе были решены следующие задачи:

- что такое большие данные;
- рассмотрены перспективы и использование больших данных в здравоохранении;
- рассмотрены методики анализа больших данных;
- рассмотрены основные задачи, стоящие перед разработчиками технологий Big Data в медицине;
- на основании представленных теоретических данных и информации о больших данных рссмотрены технологии, используемые для сбора и обработки «Больших Данных»;
- представлены возможные применения отдельных направлений технологий Big Data в биомедицине и здравоохранении;
- предствлена теоретическая часть нейронной сети.

Рассмотрим основные теоретические аспекты работы.

Что такое «большие данные»?

Однозначного определения понятия «большие данные» не существует, однако можно сослаться на два описания сути этой концепции, с которой согласится большинство людей. Первое определение предложил Мерв Адриан из компании Gartner в 2011 году: «Большие данные» — это данные, сбор, управление и обработку которых невозможно осуществить с помощью наиболее часто используемых аппаратных сред и программных инструментов в течение допустимого для пользователя времени». Другое хорошее определение появилось в докладе McKinsey Global Institute в мае 2011 года: «Большие данные» — это наборы данных, размеры которых выходят за пределы возможностей по сбору, хранению, управлению и анализу, присущих обычному программному обеспечению базы данных». Из этих определений следует, что то, что считается большими данными, будет изменяться по мере развития

технологий. То, что когда-то было «большими данными», или то, что считается «большими данными» сегодня, будет отличаться от «больших данных» завтрашнего дня. Некоторых настораживает этот аспект понятия больших данных. Приведенные определения подразумевают, что суть больших данных может отличаться в зависимости от отрасли или даже организации, если существует значительная разница в возможностях инструментов и технологий.

За $600 сегодня можно купить диск, способный вместить всю музыку мира. Каждый месяц через сеть Facebook пользователи обмениваются 30 миллиардами фрагментов информации. В среднем компании пятнадцати из семнадцати отраслей промышленности Соеди ненных Штатов имеют больше информации, чем Библиотека Конгресса США. Gartner — исследовательская и консалтинговая компания, специализирующаяся на рынках информационных технологий.

Слово «большие» характеризует не только объем, хотя понятие «большие данные» подразумевает наличие большого количества данных, оно не относится только к объему данных. Большие данные характеризуются возросшей скоростью их передачи, сложностью и разнообразием по сравнению с источниками данных прошлого.

Понятие «большие данные» подразумевает не только возросший объем, но и возросшая скорость передачи и разнообразие источников. Такие факторы, разумеется, усложняют работу с большими данными, поскольку вам приходится иметь дело не просто с большим количеством данных, а с тем, что они поступают к вам очень быстро, в сложных формах и из разнообразных источников. Методы, процессы и системы анализа, внедренные в организациях, будут использоваться до предела, а возможно, и сверх предела. Необходимо разработать дополнительные методы и процессы анализа на базе обновленных технологий и методов для того, чтобы эффективно анализировать большие данные и действовать на основании полученных результатов.

Перспективы и использование больших данных в здравоохранении

На рубеже 2012-2013 гг. технологии Big Data вышли за рамки предметной области ИТ и стали все глубже проникать в структуры управления, бизнес, промышленность и науку. Аналитики прогнозируют стремительный рост рынка инструментов и методов Big Data. По оценкам International Data Corporation (IDC), объемы хранящихся данных будут ежегодно увеличиваться на 40%, а рынок технологий и услуг Big Data в 2017 г. достиг $32,4 млрд долл. а к 2020 г. достигнет 68,7 млрд долл. Еще более оптимистичнее прогнозы объема рынка Big Data приведены в маркетинговом исследовании компании Wikibon. Согласно ее 9 прогнозу, объем рынка Big Data достиг к концу 2017 г. 50 млрд долл. Результаты специально проведенного аналитического опроса, целью которого была оценка степени внедрения технологий Big Data в различных отраслях, демонстрируют, что в системах здравоохранения различных стран мира практическое применение этих технологий пока крайне ограничено. Тем не менее целесообразность и перспективность использования технологий Big Data в медицине и системе здравоохранения в последние годы широко обсуждается профессиональным сообществом. Мировым лидером по разработке и внедрению технологий Big Data в здравоохранении на сегодняшний день являются США. Главное основание для их развития экономическая эффективность от их внедрения. По мнению аналитиков McKinsey Global Institute, использование технологий Big Data в здравоохранении США будет формировать финансовый поток объемом 300 млрд долл. в стоимостном выражении, из которых две трети за счет снижения расходов системы

здравоохранения США. Некоторые эксперты утверждают, что даже сравнительно небольшие инвестиции в массовое внедрение технологий Big Data в этой области могут в короткие сроки существенно повысить уровень качества жизни людей. Например, исследователи Калифорнийского университета (США) показали, что простой анализ данных, публикуемых в социальных се-

тях, позволяет предсказывать всплески поведения, провоцирующего ВИЧ, что дает возможность разработать систему противоэпидемических мероприятий в конкретном регионе мира. В Стратегии развития отрасли информационных технологий в Российской Федерации на 2014-2020 годы и на перспективу до 2025 года технологии обработки «Больших данных» обозначены в числе «прорывных для мировой индустрии, в которых в перспективе 10-15 лет с высокой вероятностью может быть обеспечена глобальная технологическая конкурентоспособность России». О необходимости «внедрения технологий масштабирования баз знаний и внедрения систем поддержки принятия врачебных решений в повседневную деятельность» говорилось и в государственной программе РФ «Развитие здраво- охранения» 2014 г.

Задачи, стоящие перед разработчиками технологий Big Data в медицине

Очевидно, что применение технологий Big Data для анализа все более 10 сложных массивов медицинских данных открывает новые возможности в об- ласти здравоохранения. Основные задачи, стоящие перед разработчиками технологий Big Data в медицине, определяются, главным образом, особенностями циркулирующих в современном здравоохранении и биомедицине данных. Эти данные зачастую являются непреодолимыми для обработки с помощью традиционного программного обеспечения не только из-за их объема, но и из-за разнообразия типов данных и скорости, с которой они должны анализироваться. Формирующийся из разнообразных по структуре, формату, достоверности источников массив медицинской информации, как полагают эксперты, на 78% представляет собой неструктурированный набор файлов, таблиц, рисун- ков, графиков, их описаний и зачастую противоречивых выводов и суждений. Источники медицинских данных включают в себя:

1. клинические данные для поддержки принятия решений различной спе- циализации (диагностическая, прогностическая, с элементами искусственного интеллекта, управления, уход за больными и т.д.), в виде стандартизированных данных из электронных историй болезни.

2. зарегистрированные данные с датчиков мониторинга и записывающих устройств.

3. генерируемые экспертами конкретные показатели, письменные заметки и медицинские рецепты.

4. звукозаписи и визуальные образы.

5. данные специализированных исследований.

6. данные о лекарственных препаратах.

7. данные неотложной помощи.

8. административно - паспортные данные.

9.данные о страховании и медицинском страховании.

10. социальные публикации в СМИ, в том числе Twitter - каналы, блоги, обновления статуса на Facebook и других платформ и веб - страниц.

11. данные об опыте и результатах использования методов нетрадиционной 11 медицины и непрофессиональных инициатив в области здравоохране- ния и медицины.

12. нормативные и законодательные документы из области социальной ме- дицины, общественного здравоохранения, рынка здравоохранения, политики и культуры. 13. данные медицинской науки.

Технологии, используемые для сбора и обработки «Больших Данных», можно разделить на 3 группы:

1. Программное обеспечение;

2. Оборудование;

3. Сервисные услуги.

К наиболее распространенным подходам обработки традиционных и больших данных (ПО) относятся:

SQL - язык структурированных запросов, позволяющий работать с базами данных. С помощью SQL можно создавать и модифицировать данные, а

управлением массива данных занимается соответствующая система управления базами данных.

NoSQL - термин расшифровывается как Not Only SQL (не только SQL). Включает в себя ряд подходов, направленных на реализацию базы данных, имеющих отличия от моделей, используемых в традиционных, реляционных СУБД. Их удобно использовать при постоянно меняющейся структуре данных. Например, для сбора и хранения информации в социальных сетях.

MapReduce - модель распределения вычислений. Используется для параллельных вычислений над очень большими наборами данных (петабайты* и более). В программном интерфейсе не данные передаются на обработку программе, а программа "- данным. Таким образом, запрос представляет собой отдельную программу. Принцип работы заключается в последовательной обработке данных двумя методами Map и Reduce. Map выбирает

предварительные данные, Reduce агрегирует их.

Hadoop - используется для реализации поисковых и контекстных механизмов высоконагруженных сайтов "--- Facebook, eBay, Amazon …

Отличительной особенностью является то, что система защищена от выхода из строя любого из узлов кластера, так как каждый блок имеет, как минимум, одну копию данных на другом узле. Считается одной из основополагающих технологий «больших данных». Вокруг Hadoop образовалась целая экосистема из связанных проектов и технологий, многие из которых развивались изначально в рамках проекта, а впоследствии стали самостоятельными. Со второй половины 2000-х годов идёт процесс активной коммерциализации технологии

SAP HANA - высокопроизводительная NewSQL платформа для хранения и обработки данных. Обеспечивает высокую скорость обработки запросов. Еще одним отличительным признаком является то, что SAP HANA упрощает

системный ландшафт, уменьшая затраты на поддержку аналитических систем.

Оборудование инфраструктурное - относят средства ускорения платформ, источники бесперебойного питания, комплекты серверных консолей и др.

Сервисные услуги - Серверы включают в себя хранилища данных. Сервисные услуги включают в себя услуги по построению архитектуры системы базы данных, обустройству и оптимизации инфраструктуры и обеспечению безопасности хранения данных.

Программное обеспечение, оборудование, а также сервисные услуги

вместе образуют комплексные платформы для хранения и анализа данных. Такие компании, как Microsoft, HP, EMC предлагают услуги по разработке, развертыванию решений Больших Данных и управления ими.

Возможные применения отдельных направлений технологий Big Data в биомедицине и здравоохранении

Наиболее остро необходимость новых программно - технических средств, опирающихся на методы анализа больших объемов данных, наблюдается в биоинформатике и биомедицине.

Методы полного геномного секвенирования генерируют такой большой объем данных, содержащих информацию об отдельных участках генома, что проблемой становится не только их обработка, но и запись на информационный носитель и передача копии данных в другую лабораторию. Традиционные алгоритмы анализа данных не справляются с поставленными перед ними задачами.

Ожидается, что прогресс в междисциплинарной области, объединяющей вычислительные и геномные технологии, приведет к беспрецедентным дости-

жениям персонифицированной медицины. Появление методов секвенирования высокой пропускной способности уже позволило исследователям изучить генетические маркеры на широком спектре нозологий и повысить точность и специфичность анализов более чем на пять порядков с тех пор, как было завершено секвенирование генома человека, ассоциировать генетические причины с фенотипом заболевания.

Другим направлением биомедициины, развитие которого невозможно без применения подходов и технологий Big Data, является исследование микробиома. В США проект по исследованию микробиома человека «Human Microbiome Project» был запущен одновременно с известным проектом по исследованию генома человека Human Genome Project. В ходе его реализации в рамках Национальных институтов здоровья США создан специальный центр Data Analysis and Coordination Center. Реализуется совместный китайско - европейский проект MetaHit, где ведутся активные исследования в этом направлении. В России в ряде проектов по исследованию микробиома участвует Центр исследований и разработок EMC.

Еще одной важной задачей из области биоинформатики, решить которую позволят подходы и технологии Big Data является создание и сопровождение баз данных и знаний, таких как специализированные базы белковых структур, нуклеотидных последовательностей генов, метаболических путей, клеточных ансамблей и т.п. Число и объём информации подобных баз данных стремительно растет, работа с такими огромными массивами информации требует принципиально новых подходов к обработке данных и соответствующего программного обеспечения

Нейронные сети

Нейронные сети - основные понятия и определения

В основу искусственных нейронных сетей положены следующие черты живых нейронных сетей, позволяющие им хорошо справляться с нерегулярными задачами: - простой обрабатывающий элемент - нейрон; - очень большое число нейронов участвует в обработке информации; - один нейрон связан с большим числом других нейронов (глобальные связи) ; - изменяющиеся по весу связи между нейронами; - массированная параллельность обработки информации.

Модели нейронных сетей

- Модель Маккалоха

- Модель Розенблата

- Модель Хопфилда

Задачи, решаемые на основе нейронных сетей

В литературе встречается значительное число признаков, которыми должна обладать задача, чтобы применение НС было оправдано и НС могла бы ее решить: - отсутствует алгоритм или не известны принципы решения задач, но накоплено достаточное число примеров; - проблема характеризуется большими объемами входной информации; - данные неполны или избыточны, зашумлены, частично противоречивы.

Таким образом, НС хорошо подходят для распознавания образов и решения задач классификации, оптимизации и прогнозирования. Ниже приведен перечень возможных промышленных применений нейронных сетей, на базе которых либо уже созданы коммерческие продукты, либо реализованы демонстрационные прототипы.

ЗАКЛЮЧЕНИЕ

Цель данной работы заключалась в рассмотрении применения «больших данных» в здравоохранении.

В ходе работы были рассмотрены как общие понятия «больших данных», так и их применения в здравоохранении.

Непрерывный рост объемов данных и появление более мощных аналитических инструментов дают возможность делать прогнозы об эффективности различных методов лечения пациентов с определенными характеристиками.

Сегодня мы стоим на пороге настоящего прорыва в доказательном клиническом обслуживании, так как результаты анализа больших данных станут доступны прямо в кабинете вашего лечащего врача. Решая, какой из двух-трех методов лечения выбрать, доктор теперь имеет возможность просмотреть историю ваших обращений к врачу, медицинские снимки, результаты прежних и текущих лабораторных обследований, — и все это в сопровождении рекомендаций, основанных на объективных данных. Причем все это происходит с одним пациентом на приеме у одного врача в один значимый для лечения момент. Мы, технологи, называем это mash-up (В дословном переводе — «смешение». В ИТ-индустрии mash-up — это сервис, который полностью или частично использует в качестве источников информации другие сервисы, предоставляя пользователю новую функциональность для работы. Такой сервис тоже может стать новым источником информации для других mash-up сервисов. В результате образуется сеть зависимых друг от друга сервисов, интегрированных друг с другом).

Большие данные и новые инструменты, которые необходимы для их анализа, меняют правила игры в здравоохранении. Они позволяют находить информацию о результатах лечения в настолько крупных хранилищах данных о пациентах, что прежде проанализировать их было невозможно — по крайней мере, в реальном времени. Открывающийся таким образом новый взгляд на вещи дает врачам клиническое средство доказательной медицины,

когда учитывается полная информация о вашем здоровье за все годы, собранная из множества независимых источников. Это могут быть и результаты лабораторных исследований, и записи терапевтов, и данные о стационарном лечении, и история обращений за прошлые годы. Когда такой инструмент в нужный момент оказывается в распоряжении врача — это и есть изменение правил игры.