Системный анализ в медицине и биологии

Методы и средства комплексного интеллектуального анализа медицинских данных*

А. А. Баранов, Л. С. Намазова-Баранова, И. В. Смирнов, Д. А. Девяткин, А. О. Шелманов, Е. А. Вишнева, Е. В. Антонова, В. И. Смирнов, А. В. Латышев

Аннотация. Выполнен обзор методов и систем интеллектуального анализа медицинских данных и клинических текстов на естественном языке. Проанализирован типовой состав данных многопрофильного педиатрического центра и выявлены направления применения и задачи комплексного интеллектуального анализа медицинских данных. Предложена архитектура системы комплексного интеллектуального анализа медицинских данных, а также выбраны платформы для ее реализации.

Ключевые слова: *интеллектуальный анализ медицинских данных, автоматическая обработка медицинских текстов, медицинская информационная система, большие данные, grid-системы.*

Введение

Ведущие медицинские организации активно внедряют системы поддержки принятия решений, которые, используя методы интеллектуальной обработки данных, помогают специалистам в задачах постановки диагнозов, назначения курса лечения, прогнозирования развития заболеваний. На вход систем поддержки принятия решений поступает информация из систем ведения электронных историй болезни, которые аккумулируют большие объемы разнородной информации, генерируемой медицинской организацией: показатели здоровья пациентов, результаты обследований, данные о проведении лечебных процедур и др. При этом в одной организации может быть сразу несколько медицинских систем, данные в которых хранятся в разных форматах, соответствующих разным стандартам. Каждая такая система, как правило, предназначена для решения узкого круга задач, например, для лечения определенного заболевания или проведения определенной диагностики [41]. На практике возникает необходимость в системе комплексного интеллектуального анализа данных, которая могла бы агрегировать и анализировать разнотипную информацию, поступающую от всех медицинских систем организации. Из-за больших объемов анализируемой информации подобная комплексная система должна использовать технологии работы с большими данными.

Помимо структурированной информации медицинские организации генерируют огромный объем неструктурированной информации, которая содержится в текстах на естественном языке (ЕЯ). Большинство историй болезни, анамнезов, эпикризов,

* Работа выполнена при поддержке РФФИ (проект № 13–04–12062).

а также отчетов о проведении клинических мероприятий: операций, анализов и обследований, таких как рентгеновские, ультразвуковые исследования, записываются в виде текстов на ЕЯ. Эти тексты содержат много полезной информации, которую необходимо извлечь и структурировать. В области обработки текстов на ЕЯ выделилось отдельное актуальное быстроразвивающееся научное направление, которое занимается проблемой анализа клинических текстов. В рамках этого направления разрабатываются специализированные системы, решающие задачи извлечения информации из клинических текстов и ее структурирования. Информация, полученная из текстов, может существенно обогатить базы знаний и данных, на основе которых работают медицинские системы поддержки принятия решений, что, в конечном счете, может повысить их эффективность. Большинство существующих методов и систем анализа медицинских текстов работают только с английским языком, системы анализа медицинских текстов на русском языке отсутствуют.

Таким образом, возникают две актуальные проблемы: проблема создания системы комплексного интеллектуального анализа медицинских данных, которая могла бы агрегировать и анализировать получаемую из разнородных источников разнотипную информацию, включая числовые, графические и текстовые данные, а также проблема разработки методов и инструментов интеллектуального анализа клинических текстов на русском языке.

В статье рассматриваются вопросы создания системы комплексного интеллектуального анализа медицинских данных на примере многопрофильного педиатрического центра. В первой части статьи проведен обзор методов и систем интеллектуального анализа медицинских данных и текстов. Во второй и третьей частях статьи предложены решения по созданию системы комплексного интеллектуального анализа медицинских данных (на примере многопрофильного педиатрического центра), приведен типовой состав медицинских данных, сформулированы требования к системе, предложена архитектура системы комплексного интеллектуального анализа медицинских данных и выбраны платформы для анализа текстов и данных в рамках этой архитектуры.

1. Методы и системы интеллектуального анализа медицинских данных и текстов

В этом разделе представлен обзор методов и систем интеллектуального анализа медицинских данных и текстов.

1.1. Методы и системы интеллектуального анализа медицинских данных

Методы интеллектуального анализа данных, применяющиеся в медицине, можно разделить на несколько групп в соответствии с решаемыми с помощью них задачами (табл. 1): прогнозирование течения болезни, воздействия препарата или группы препаратов, уровня смертности; обследование — постановка диагноза на основе совокупности симптомов; классификация — уточнение диагноза; поиск ассоциаций — поиск скрытых зависимостей между различными показателями здоровья пациентов [36]. Рассмотрим далее основные методы интеллектуального анализа данных, применяемые для обработки медицинской информации.

Таблица 1

Задачи интеллектуального анализа данных в медицине и методы, применяемые для их решения

Цель анализа	Методы с учителем	Методы без учителя
Прогнозирование	Метод наименьших квадратов Логистическая регрессия Нейронные сети Деревья принятия решений SVM Сплайны	–
Обследование	Деревья принятия решений	Метод главных компонент Кластеризация Анализ ссылок
Классификация	Деревья принятия решений Нейронные сети Дискриминантный анализ Бустинг Наивный Байесовский классификатор	Кластеризация Самоорганизующиеся карты Кохонена
Поиск ассоциаций	–	Факторный анализ Априорный алгоритм

Многие задачи в биоинформатике и медицине решаются с помощью методов классификации. Например, с помощью методов классификации массивов информации о ДНК клеток, измененных при различных заболеваниях, повышается точность диагностики схожих заболеваний. В работе [25] был найден способ дифференцирования миелоидной лейкемии и острой лимфобластной лейкемии. Пра-

вильная диагностика этих заболеваний важна для назначения эффективного курса лечения. Другим практическим применением методов классификации в медицине является определение вида опухоли груди (доброкачественная или злокачественная), а также предсказание вероятности рецидивов опухоли. В работе [38] было проведено исследования точности работы 56 алгоритмов классификации, включая бустинг, из программного пакета Weka [53] на двух наборах диагностических данных пациентов с опухолями груди. В результате проведенных экспериментов автору не удалось выявить метод, который бы одинаково хорошо работал на обоих наборах данных.

В работах [26, 27] оценивалась эффективность методов классификации на различных наборах медицинских данных из репозитория Kent Ridge Biomedical Dataset Repository [30]. Эти исследования показали, что точность классификации зависит от набора анализируемых данных и применяемого метода извлечения и ранжирования ключевых признаков. Согласно [55], деревья принятия решений являются наиболее распространенным алгоритмом классификации в системах интеллектуальной обработки медицинских данных благодаря возможности отображения процесса принятия решения в понятной эксперту форме. Алгоритмы, такие как SVM, применяется реже, поскольку процесс приятия решений в нем не прозрачен для эксперта.

Методы и модели прогнозирования в медицине обычно используются для вычисления вероятности появления изменений в состоянии больного. Например, в работе [16] были построены три модели прогнозирования выживаемости при раке груди: на основе нейронных сетей, деревьев принятия решений и логистической регрессии. Больному присваивался класс «выживший», если пациент был жив более 5 лет после постановки диагноза. Было произведено сравнение точности этих моделей. Деревья принятия решений показали лучшую точность (93,6 %), тогда как логистическая регрессия показала наихудший результат (89,2 %).

В работе [45] был предложен метод прогнозирования выживаемости пациентов с терминальной стадией хронической почечной недостаточности. Наблюдение за такими пациентами в среднем составляет около 3 лет. На выживаемость таких пациентов влияет множество факторов. Для выявления наиболее важных факторов, использовались теория приближенных множеств и деревья принятия решений. При помощи обоих методов строились правила, из которых извлекались значимые факторы. Авторы выявили следующие медицинские факторы, влияющие на выживаемость: диагноз, время на диализе, отклонения от целевого веса, артериальное давление, уровни кальция и калия, общий объем крови.

Другим примером практического применения методов прогнозирования в медицине является априорное определение стоимости лечения пациента. В работе [8] представлен метод прогнозирования стоимости лечения. Предварительно было задано 5 категорий стоимости лечения. Для прогнозирования использовался метод на основе деревьев принятия решений. В качестве экспериментальных и тестовых данных использовалась информация, собранная страховыми агентствами за трехлетний период. Данные первых двух лет использовались для обучения классификатора, а данные за третий год — для тестирования. Авторы применяли различные стратегии извлечения признаков. Максимальная точность составила 84,6 %.

Методы поиска ассоциаций используются для обнаружения скрытых зависимостей между признаками. Априорный алгоритм [3] позволяет находить ассоциации в очень больших массивах данных. Например, в работе [11] статистические методы, деревья принятия решений и методы поиска ассоциаций были успешно применены для создания правил лечения гипертонии. Другим примером практического использования методов поиска ассоциаций стало определение комбинаций совместно употребляемых лекарственных препаратов. В работе [12] исследовалось то, как часто антациды прописываются с другими лекарственными препаратами и какими именно. Антациды — наиболее широко применяющиеся лекарства для облегчения симптомов изжоги при язвенной болезни желудка и гастрите путем нейтрализации желудочной кислоты. В этой работе при помощи методов поиска ассоциаций были построены 36 правил, которые указывают наиболее часто прописываемые совместно с антацидами наборы лекарств.

Методы кластеризации часто используются для анализа массивов генетической информации. В работе [52] была выполнена кластеризация массива, содержащего ДНК 86 видов опухолей груди. Было получено два кластера. В первый кластер вошли опухоли, дающие рецидив в 34 % случаев, во второй — в 70 %. Первый кластер условно можно назвать «плохо прогнозируемыми опухолями», а второй — «хорошо прогнозируемыми опухолями». Далее эта информация использовалась для повышения точности прогнозирования развития опухолей.

Рассмотренные методы интеллектуальной обработки данных используются в системах поддержки принятия решений, которые активно внедряются в ведущие медицинские клиники. Эти системы помогают врачам в задачах постановки диагнозов, назначения курса лечения, прогнозирования развития заболеваний. Примером может служить система контроля больничных инфекций, разработанная в университете Алабамы [9]. Для раннего распознава-

ния и лечения вспышки нозокомиальной инфекции необходимо постоянное активное наблюдение за больными. Созданная авторами система наблюдения автоматически определяет новые шаблоны в данных о распространении инфекции. Эта система использует ассоциативные правила и данные об уходе за пациентом, полученные из лабораторных систем управления информацией, и генерирует шаблоны, которые уточняются специалистом по контролю над инфекцией. Разработчики системы делают вывод, что расширенный контроль над инфекциями с помощью системы интеллектуальной обработки данных более чувствителен, чем традиционные наблюдения за распространением инфекции.

Несколько приложений поддержки принятия решений, имеющих отношение к здравоохранению, было разработано в IBM. Эти приложения базируются на компьютерной платформе IBM Watson [56]. Одно из приложений предназначено для рекомендации методов лечения раковых заболеваний, другие приложения — для рассмотрения и утверждения процедур, связанных со страхованием здоровья. Система рекомендации методов лечения раковых заболеваний была разработана совместно с институтом Memorial Sloan-Kettering (MCK) в Нью-Йорке, специализирующемся на лечении рака. IBM Watson анализирует как структурированные данные, так и тексты на естественном языке. Система автоматически извлекает информацию из медицинских заключений, медицинских журналов и клинических испытаний в области онкологии.

Система Manteia позволяет предсказывать генетически-обусловленные заболевания человека [49]. В ней реализованы модели позвоночных арготизмов, которые генерируют большие объемы данных. В Manteia сохраняются данные, порожденные для четырех видов позвоночных: человек, мышь, курица и рыба данио. Информация, сохраненная в системе, покрывает различные аспекты развития эмбриона и генетические расстройства, приводящие к ненормальным фенотипам и генетическим болезням в моделях животных. Система включает в себя ряд инструментов интеллектуальной обработки данных, которые позволяют анализировать и аннотировать эту информацию.

Интеллектуальный анализ медицинских данных применяется в системах поддержки принятия решений при лечении опухолей головного мозга [41], а также при изучении склерозов. Система [23] позволяет извлекать скрытые зависимости между различными иммунологическими маркерами, используя нечеткую кластеризацию, эволюционное программирование (PST) и сеть семантических связей (AutoCM). Эта интеллектуальная адаптивная система позволяет задавать степень ассоциации каждой переменной со всеми другими и отображать карту основных связей переменной. Матрица связей, визуализированных в виде минимальных остовных деревьев при помощи AutoCM, определяет нелинейные ассоциации между переменными и определяет схемы соединения кластеров. С помощью этой системы были обнаружены ранее неизвестные связи между большим набором фактов, связанных со склерозом.

Общей особенностью рассмотренных систем является их узкая направленность — каждая из них создана для решения одной задачи при помощи небольшого набора методов интеллектуального анализа данных. Целью нашей работы является создание системы комплексного интеллектуального анализа данных, решающей широкий спектр задач, поэтому, в архитектуре такой системы должна быть предусмотрена возможность использования различных методов интеллектуального анализа данных и динамического комбинирования этих методов.

1.2. Системы анализа клинических текстов

В области понимания текста (text understanding) существует множество методов и программных средств, позволяющих структурировать тексты на естественном языке и извлекать из них информацию. Большинство существующих решений ориентированы на обработку текстов общего характера, например, таких как новостные сообщения. Однако стилистика клинических текстов сильно отличается от стилистики обычных текстов, поэтому требуется как значительная доработка существующих методов и инструментов по анализу текстов на ЕЯ, так и создание новых специфичных подходов. Богатая медицинская терминология также предполагает разработку объемных лингвистических ресурсов: номенклатур, кодификаторов и тезаурусов. Эти особенности позволяют выделить анализ клинических текстов в обособленное направление исследований в области обработки ЕЯ.

Последние десять лет наблюдается возрастающий интерес исследователей к проблемам анализа клинических текстов. Большое внимание к этой области со стороны академического сообщества привело к проведению ряда семинаров и соревнований по задачам извлечения информации из клинических текстов и поиска медицинских данных, например, 2010 i2b2 /VA Challenge [51], 2012 i2b2 /VA Chal¬lenge [48], CLEF eHEALTH 2013 [39], CLEF eHEALTH 2014 [35], SemEval 2014 Task 7 [43], SemEval 2015 Task 6, 14 [44]. Среди этих задач — расшифровка медицинских аббревиатур и сокращений; выделение концептов, обозначающих заболевания, патологии, вмешательства, медицинские препараты, обследования; выявление семантических связей между этими концептами, а также определение атрибутов этих концептов.

Исследования в этой области привели к разработке ряда прикладных систем и платформ, специализирующихся на комплексном компьютерном лингвистическом анализе медицинских текстов, некоторые из которых уже применяются в клиниках для повышения качества медицинских услуг.

Большие усилия направлены на стандартизацию и унификацию клинических записей, и в этом помогают методы извлечения информации из текстов на естественном языке. Система UMLS [31], разработанная в Национальной медицинской библиотеке США, содержит метатезаурус, словари, семантическую сеть и программные компоненты, которые позволяют сопоставлять концепты из разных медицинских и биомедицинских баз знаний друг с другом и находить их в текстах на естественном языке. Таким образом, эта система помогает преодолеть две проблемы информационного поиска: вариативность написания терминов в разных источниках и разрозненность информации, представленной в отдельных медицинских и биомедицинских базах. UMLS интегрирует в себе более 100 баз знаний, среди которых МКБ-10 [29], MeSH [34], SNOMED-CT [46], LOINC [28], DSM [4], Gene Ontology [7]. При этом многие ресурсы имеют переводы на несколько языков. Из русскоязычных медицинских ресурсов в UMLS на данный момент интегрированы MeSH [1] и МКБ-10 [2]. Среди программных компонентов UMLS стоит отметить инструмент MetaMap [6] — анализатор, который позволяет искать медицинские термины в текстах на ЕЯ и сопоставлять их с концептами метатезауруса. MetaMap и другие компоненты системы UMLS используются во многих других системах обработки клинических текстов и в системах поиска медицинской и биомедицинской информации.

Одной из первых систем, примененных на практике для решения прикладных задач в области обработки клинических текстов, была проприетарная система Medical Language Extraction and Encoding System (MedLEE) [21, 22]. Она изначально была разработана для обработки отчетов о рентгеновских исследованиях грудной клетки пациентов в медицинском центре Columbia Presbyterian Medical Center (CPMC), а затем адаптирована к другим областям. В частности, она также применялась для анализа эпикризов и отчетов по маммографическим, эхокардиологическим и другим исследованиям. Основной задачей системы MedLEE являлось извлечение, структурирование и кодирование клинической информации, содержащейся в текстах медицинских отчетов. Используя морфологический, синтаксический и другие виды анализа, MedLEE заполняла заданный шаблон клинического исследования и заносила структурированную информацию в базу данных, которая затем использовалась другими информационными системами и системами поддержки принятия решения медицинского центра.

Система Health Information Text Extraction (HITEx) [57] — это адаптируемая система для обработки медицинских текстов с открытым исходным кодом, разработанная в National Center for Biomedical Com¬puting в рамках масштабного проекта Informatics for Integrating Biology & the Bedside (i2b2). В основе HITEx лежит платформа GATE [15]. Система предоставляет набор модулей, которые ориентированы на обработку клинических текстов. Среди функций этих модулей присутствуют: поиск и классификация разделов клинических текстов; токенизация и выделение предложений; морфологический анализ (с разрешением омонимии); поверхностный синтаксический анализ; определение отрицаний; выявление медицинских терминов и определение их кодов в метатезаурусе UMLS; выделение n-грамм; поиск составных медицинских концептов; поиск предложений, указывающих на никотиновую зависимость пациента. Система применялась для поиска в архивах клинических текстов признаков, указывающих на наличие астмы и других хронических заболеваний легких у пациентов. Тестирование, проводившееся совместно с экспертами клиники Brigham and Women's Hospital, показало, что система HITEx позволяет определять наличие бронхолегочных заболеваний у пациентов точнее, чем при использовании только лишь кодов МКБ-9, проставленных в историях болезни работниками клиники вручную. Система также использовалась для определения уровня никотиновой зависимости пациентов, информация о котором не была закодирована МКБ-9 в клинических текстах. Тестирование выявило высокую точность решения этой задачи системой HITEx, что показало перспективность применения методов автоматического извлечения структурированной информации о пациентах из клинических текстов на ЕЯ.

Одной из наиболее перспективных и быстроразвивающихся платформ по анализу медицинских и биомедицинских текстов является система cTAKES [42]. Это модульная система с открытым исходным кодом, которая разрабатывается сообществом исследователей из разных институтов совместно с частной клиникой Mayo [33]. Основное предназначение cTAKES заключается в создании основы для систем извлечения информации из клинических текстов. Система реализована на платформе UIMA [17] с применением фреймворка OpenNLP [5]. cTAKES предоставляет набор программных модулей, которые используют наборы правил и модели машинного обучения, настроенные на анализ клинических текстов. Система позволяет проводить: 1) базовую обработку медицинских текстов, включая морфологический анализ и различные виды синтаксического

анализа; 2) извлечение именованных сущностей, определение их типа и характера упоминания их в тексте; 3) сопоставление терминов с концептами метатезауруса UMLS; 4) поверхностный семантический анализ (установление семантических ролей); 5) разрешение кореференции; 6) извлечение семантических связей между сущностями. В системе также присутствуют готовые модули для решения прикладных задач: извлечения названий лекарственных препаратов и определения статуса пациента курящий/некурящий. Платформа интегрирует в себе большое число современных разработок в области анализа медицинских и биомедицинских текстов и продолжает развиваться. Заметим, что cTAKES ориентирована на обработку только английских текстов и на сегодняшний день не поддерживает другие языки.

Стоит также отметить ряд других систем анализа медицинских текстов. Системы SPRUS/SymText/MPLUS [13] использовались для анализа отчетов по снимкам легких и выявления пневмонии. Проприетарные системы IBM BioTEKS [32] и IBM MedKAT [14] интегрируют разработки компании IBM в области анализа текстов по биомедицине и клинических текстов соответственно. Система MedEx [54] применялась для поиска в тексте названий лекарств и дозировок. KMCI — инструмент для поиска в клинических текстах концептов из метатезауруса UMLS и их кодирования. IBM Watson Oncology [56] — проприетарная система, ориентированная на решение задач поддержки принятия решений в области лечения онкологических заболеваний.

Направление анализа клинических и биомедицинских текстов стремительно развивается, в нём возникает все больше подходов, методов, ресурсов и программных инструментов. Однако разработанные в рамках этого направления системы по большей части ориентированы на анализ текстов на английском языке. Многие системы основаны на известных платформах обработки текстов на ЕЯ общего назначения. Наиболее часто используемые из них — GATE и Apache UIMA.

2. Состав исходных данных и задачи системы комплексного интеллектуального анализа медицинских данных

В этом разделе описан типовой состав медицинских данных, накапливаемых лечебными учреждениями (на примере многопрофильного педиатрического центра), и определены задачи системы комплексного интеллектуального анализа медицинских данных.

2.1. Состав анализируемых данных

Определим типовой состав медицинских данных, накапливаемых лечебным учреждением (на примере многопрофильного педиатрического центра) и представляющих интерес для интеллектуального анализа. Они включают текстовые, графические и числовые данные следующих типов:

1. Лабораторные данные:
   - биохимические показатели сыворотки крови;
   - показатели кислотно-основного состава организма;
   - показатели свертывающей системы крови (коагулограмма), времени кровотечения и свертываемости;
   - определение группы крови, резус-фактора, фенотипа эритроцитов;
   - показатели клинического анализа крови;
   - иммунологические характеристики (например, иммуноглобулины сыворотки крови);
   - показатели иммунного ответа (к различным антигенам возбудителей/вирусов/паразитов);
   - наличие антител к собственным белкам и компонентам клетки;
   - определение гормонального состава сыворотки крови;
   - определение наличия/титра ферментов и кофакторов;
   - показатели активности ферментов;
   - молекулярная диагностика, секвенирование мутаций, определение полиморфизма генов;
   - общий анализ мочи;
   - биохимический анализ мочи и клеточного остатка;
   - анализ кала.
2. Инструментальные и визуализационные методы диагностики:
   - ультразвуковые исследования;
   - рентгенография, томография с контрастом и/или функциональными пробами;
   - магнитно-резонансное исследование;
   - сцинтиграфия;
   - исследование электрической активности и проводимости органов и систем — электрокардиография, электроэнцефалография, электронейромиография;
   - исследование функции внешнего дыхания: спирометрия и легочные объемы — бодиплетизмография;
   - комплексные исследования (полисомнография).
3. Показатели комплексного клинического осмотра:
   - специфические и общие жалобы;
   - подробный анамнез (история) жизни (развития) пациента и непосредственно возникновения болезни;

- наследственная предрасположенность (семейный анамнез);
- уточнение наличия вредных (отягощающих состояние) факторов окружения и внешней среды;
- осмотр органов и систем (выявление физиологических показателей и/или патологических данных, характеризующих нозологическую форму либо имеющих синдромальный характер). Анализируются:
  1) кожа и подкожная жировая клетчатка: наличие и характер высыпаний, отеки;
  2) дыхательная система: носовое дыхание, отделяемое из носовых ходов, храп, осиплость голоса, кашель (характер, время появления), мокрота, боли в груди или спине (характер, локализация, связь с дыханием, кашлем), одышка (затруднение вдоха и/или выдоха), приступы удушья, свистящее дыхание, характер перкуторного звука, аускультативная картина в легких (проводится или нет во все отделы), характеристики вдоха/выдоха, хрипы есть/нет, характер хрипов, влияние кашля на хрипы;
  3) сердечно-сосудистая система: цианоз кожных покровов, одышка, боли в области сердца, ощущение сердцебиения и «перебоев», отеки (время появления, локализация), пульс, перкуторные границы сердца, аускультативная картина (ЧСС, наличие шумов и их локализация);
  4) система органов пищеварения: наличие и характер налета на языке, глотание, тошнота, рвота, срыгивания (у младенцев), отрыжка или изжога, боли в животе (характер, локализация, иррадиация, связь с приемом пищи), характер и частота стула; пальпация живота, болезненность при пальпации вокруг пупка, по ходу толстой кишки, в точке проекции желчного пузыря;
  5) мочевыделительная система: боли в животе и в поясничной области, частота мочеиспусканий, цвет мочи, недержание мочи, отеки, есть или нет болезненность в поясничной области при поколачивании;
  6) опорно-двигательная система: боли в конечностях, мышцах, суставах (характер, локализация, связь с различными факторами — от времени суток до метеоусловий), изменение формы суставов, характер движений, наличие травм;
  7) эндокринная система: нарушение волосяного покрова, изменения кожи (чрезмерная потливость или сухость), нарушение роста и массы тела, вторичные половые признаки;
  8) нервная система и органы чувств: головные боли и головокружения, судороги, тики, нарушения со стороны органов чувств, характер рефлексов и ответа и на раздражители; очаговая симптоматика — черепно-мозговые нервы, менингеальные знаки.

Выбор каждого из указанных параметров, а также их сочетание, будет определяться для каждого конкретного случая соответствующей нозологической формой и/или патологическим синдромом.

2.2. Направления применения системы комплексного интеллектуального анализа медицинских данных

В целом, системы интеллектуальной обработки медицинских данных имеют следующие направления применения [36, 55]: прогнозирование, классификация клинических случаев (диагностика), поиск похожих клинических случаев, наблюдение за состоянием пациентов. Рассмотрим предлагаемый расширенный список направлений, составленный с учетом специфики применения системы комплексной интеллектуальной обработки данных в многопрофильном педиатрическом центре.

Одним из основных направлений применения системы является дифференциальная диагностика состояния пациента: выявление заболевания, его стадии, характера течения болезни. Необходимо предусмотреть возможность пошаговой диагностики болезни пациента с уточнением диагноза пациента на каждом шаге.

Другим важным направлением применения является прогнозирование изменения клинического состояния пациента при применении различных видов вмешательств и при отсутствии вмешательств. Под вмешательством понимаются различные диагностические или лечебные мероприятия, как медикаментозные, так и немедикаментозные, реабилитационные и профилактические, а также хирургические операции, перемещение пациента в другие лечебные учреждения, либо внутри лечебного учреждения.

Еще одним важным направлением применения системы является отслеживание опасных — критических — изменений в показателях здоровья пациента. К таким изменениям в показателях здоровья будем относить резкое ухудшение состояния пациента, вызванное течением болезни, либо реакцией на вмешательства.

Помимо основных направлений применения, система должна будет осуществлять мониторинг и автоматическую экспертизу действий медицинского персонала. В ходе экспертизы должны оцениваться: адекватность вмешательства диагнозу, корректность дозировок и длительности приема лекарственных средств, соответствие организационных действий (выписка, перевод в определенную палату, другое лечебное учреждение) состоянию больного, поиск похожих клинических случаев.

На основе состава исходных данных и предложенных типовых направлений применения были определены задачи комплексного интеллектуального анализа медицинских данных. Приведем далее наиболее значимые из них:

1) Поиск структурированных данных по запросу. Предлагается хранить структурированные данные в реляционной форме и использовать для поиска стандартные средства и методы для работы с реляционными базами данных.
2) Поиск скрытых логических и статистических закономерностей в заданных наборах медицинских данных. Для этого предлагается использовать известные статистические и логические методы интеллектуального анализа данных, а также разработать комбинированные логико-статистические методы обработки данных.
3) Классификация и предсказание признаков пациентов на основе выявленных закономерностей для решения обобщенных задач диагностики и прогнозирования.
4) Группирование структурированных данных. Группирование будет выполняться при помощи методов кластеризации.
5) Работа со сверхбольшими массивами данных, т. е. при разработке методов должна предусматриваться возможность их параллельной и распределенной реализации.
6) Лингвистический анализ текстовых документов. Для решения этой задачи будет применен метод глубокого лингвистического анализа, использующий реляционно-ситуационную модель текста.
7) Поиск похожих текстовых документов на естественном языке. Для этого будет применен метод поиска близких текстовых документов.
8) Поиск текстовых документов по запросу на естественном языке. Предлагается решать указанную задачу с помощью метода полнотекстового семантического поиска информации [37].

3. Архитектура системы комплексного интеллектуального анализа медицинских данных

В этом разделе предлагается архитектура системы комплексного интеллектуального анализа медицинских данных и обсуждаются вопросы её реализации на основе современных программных платформ и технологий анализа больших данных.

3.1. Выбор архитектуры вычислительных систем

Система интеллектуального анализа данных может быть создана на основе различных архитектур вычислительных систем: многопроцессорные вычислительные системы, кластеры, вычисления в интернете и грид-приложения. В связи со спецификой решаемой задачи к вычислительной системе предъявляются следующие требования:

1) Небольшая стоимость. Большинство лечебных учреждений имеет ограниченный бюджет на развитие ИТ-инфраструктуры.
2) Ресурсы (базы данных) системы могут быть значительно разнесены друг с другом географически.
3) Должна быть обеспечена защищенность каналов передачи данных и узлов обработки информации от взлома с целью предотвращения разглашения врачебной тайны и пересылки заведомо искаженных данных.
4) Большой объем разнородных анализируемых данных и обрабатываемых запросов.
5) Должна обеспечиваться масштабируемость системы, на случай расширения медицинской организации либо увеличения числа ресурсов, доступных для анализа.

Определим, насколько существующие архитектуры вычислительных систем соответствуют сформулированным выше требованиям.

Многопроцессорные вычислительные системы характеризуются высокой стоимостью приобретения и обслуживания, но при этом позволяют собрать в одном месте всю вычислительную инфраструктуру, что упрощает мероприятия по обеспечению защищенности данных. Однако такие системы практически не масштабируются [40].

Кластер представляет собой набор независимых вычислительных машин, связанных с помощью внутренней сети и управляющихся единой системой. Кластер обеспечивает высокую производительность, доступность, балансировку нагрузки и масштабируемость. К недостаткам кластерной архитектуры следует отнести то, что она не позволяет связать между собой в единую вычислительную сеть удаленные друг от друга ресурсы различных организаций. Кроме того, кластерная архитектура часто основывается на закрытых стандартах взаимодействия узлов сети, что делает такую систему менее переносимой.

Для реализации вычислений в интернете на компьютеры пользователей, желающих участвовать в вычислениях, устанавливается специальное приложение, использующее часть ресурсов компьютера пользователя для решения фрагмента общей задачи. Такой подход позволяет для решения задач мобилизовать действительно большие, географически удаленные друг от друга, ресурсы [20]. Однако проблема обеспечения защищенности данных при вычислениях в интернет еще не решена.

В грид-приложениях все распределенные ресурсы, которые объединены в сеть, можно рассматривать как виртуальный суперкомпьютер. Грид-приложение не имеет централизованного управления, так как каждая отдельная система ему не принадлежит и управляется администратором этой системы [18]. Отметим положительные стороны этой вычислительной архитектуры:

1) Децентрализованное управление. В гриде контроль над ресурсами является децентрализованным, что позволяет использовать различные политики управления и местные системы управления.
2) Открытые технологии. Грид-приложение использует открытые протоколы и стандарты.
3) Хорошая масштабируемость. Грид-приложение можно практически неограниченно масштабировать, добавляя новые ресурсы.

В результате анализа было выявлено, что наиболее полно заданным критериям эффективности отвечают системы интеллектуального анализа данных, архитектуры которых основаны на грид. В таких системах обычно используется специализированная архитектура DataMining Grid Architecture (DMGA) [10]. Для создания системы комплексного интеллектуального анализа данных в многопрофильном педиатрическом центре предлагается использовать аналогичную архитектуру.

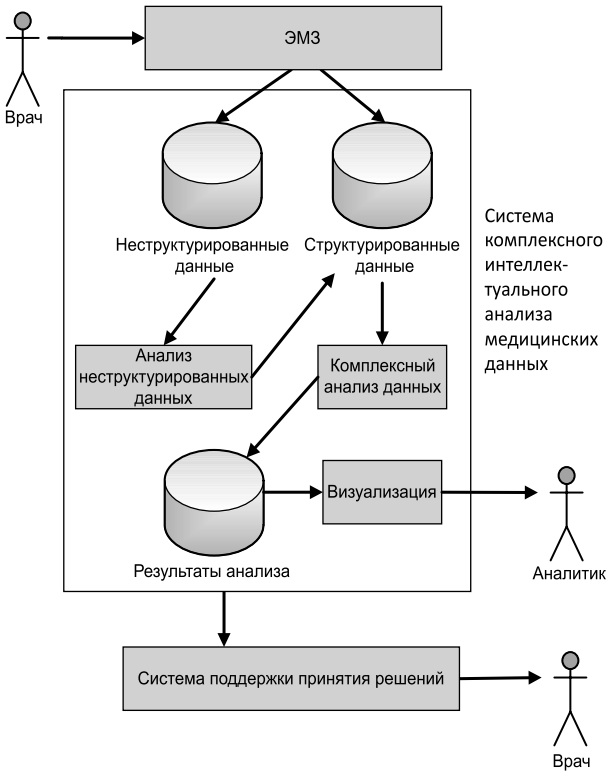

Рис. 1. Место системы комплексного интеллектуального анализа данных в информационной инфраструктуре лечебного учреждения

3.2. Логическая архитектура системы

Создаваемая система комплексного интеллектуального анализа медицинских данных должна интегрироваться с медицинскими информационными системами (МИС) и системами поддержки принятия решений в лечебных учреждениях. На рис. 1 показано взаимодействие создаваемой системы комплексной интеллектуальной обработки медицинских данных с остальными элементами информационной инфраструктуры лечебного учреждения, а также основные хранилища, блоки обработки и каналы передачи данных. Опишем далее потоки данных и работ в лечебном учреждении.

Врач заполняет истории болезней пациентов, которые сохраняются в МИС в виде электронных медицинских записей (ЭМЗ). Далее структурированные данные из ЭМЗ сохраняются в хранилище структурированных данных. Текстовая информация из ЭМЗ сначала попадает в хранилище неструктурированных данных. Далее эта информация обрабатывается блоком анализа неструктурированных данных, структурируется и записывается в хранилище структурированных данных. Затем блок комплексного анализа данных выполняет интеллектуальный анализ структурированных данных и помещает результаты в хранилище. Данные из хранилища результатов могут быть показаны аналитику. Эти результаты также могут пополнять базу знаний медицинской системы поддержки принятия решений.

Как было сказано выше, DMGA наилучшим образом удовлетворяет требованиям к архитектуре системы комплексного интеллектуального анализа медицинских данных. Архитектура DMGA основана на спецификации грид-систем OGSA [19]. Спецификация OGSA предполагает, что каждый компонент распределенной вычислительной системы представляет собой сервис. При этом, различные сервисы, составляющие единую систему могут администрироваться независимо друг от друга владельцами сервисов. В рамках спецификации OGSA выделяют следующие группы сервисов: инфраструктурные, сервисы контроля выполнения задач, сервисы данных, сервисы безопасности, сервисы управления ресурсами, информационные сервисы. Все сервисы взаимодействуют друг с другом одинаковым образом при помощи протокола SOAP.

Разрабатываемая система комплексного интеллектуального анализа медицинских данных представляет собой набор репозиториев и сервисов, работающих на удаленных друг от друга узлах вычислительной сети [47]. Помимо основных сервисов, определенных стандартом OGSA, вводится дополнительный сервис

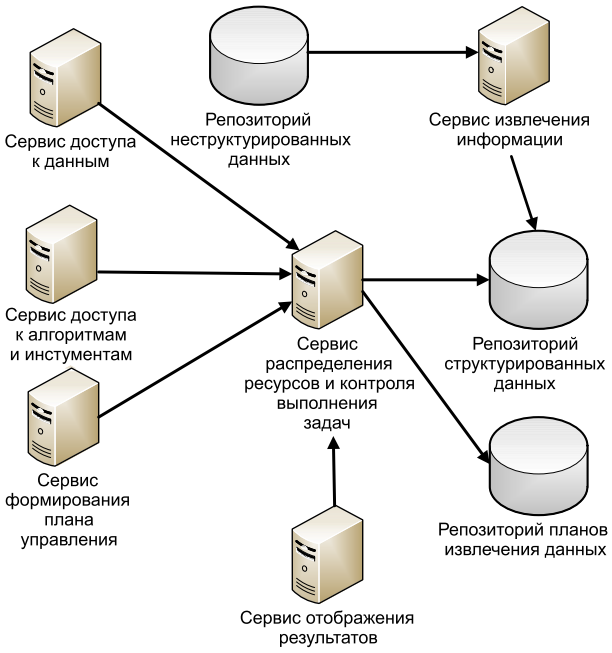

Рис. 2 Архитектура системы комплексного интеллектуального анализа медицинских данных

извлечения информации из неструктурированных данных (рис. 2).

Рассмотрим эти сервисы подробнее.

1) Сервис доступа к данным предназначен для публикации, поиска и передачи наборов данных, которые будут использоваться системой. Эти наборы располагаются в репозиториях структурированных и неструктурированных данных.
2) Сервис доступа к алгоритмам и инструментам занимается поиском инструментов и алгоритмов, которые будут использоваться в процессе интеллектуального анализа данных. Эти алгоритмы и инструменты реализуют в создаваемой системе методы классификации, кластеризации, регрессионного анализа, автоматического порождения гипотез.
3) Сервис формирования плана управления обеспечивает синтез алгоритмов решения задач системы путем создания соответствующего плана выполнения и добавления набора ограничений на ресурсы. Для каждого варианта использования создаваемой системы будет построен индивидуальный план управления. Разработанные планы управления сохраняются в репозитории планов извлечения данных.
4) Сервис отображения результатов предлагает средства для представления и визуализации извлеченной информации, а также позволяет сохранять её в подходящем формате для дальнейшего использования.
5) Сервис извлечения информации использует методы лингвистического анализа для извлечения структурированных данных из медицинских текстов. В частности, предлагается извлекать из историй болезни упоминания болезней, симптомов, диагностических процедур и медицинских вмешательств, названия лекарственных препаратов и др. Помимо этого, предлагается определять свойства найденных в тексте сущностей: например, эффект от применения лечебных мероприятий, степень тяжести и характер течения болезни, области тела, с которыми связано заболевание или расстройство функции и др. Извлеченная информация будет использована в качестве признаков при интеллектуальном анализе медицинской информации.

3.3. Программные платформы для интеллектуального анализа данных и текстов

Грид-приложения интеллектуального анализа данных обычно разрабатывают на основе готовых специализированных программных платформ. Наиболее распространенными платформами такого типа являются Globus Toolkit [24] и UNICORE [50]. В табл. 2 представлены их сравнительные характеристики.

Таблица 2

Сравнительные характеристики грид-платформ

Название платформы	Globus Toolkit	UNICORE
Лицензия	Apache	BSD
Графический клиент	Сторонние разработки	Присутствует
Тип	Набор инструментов для создания грид-приложений интеллектуального анализа данных	Грид-платформа интеллектуального анализа данных
API для клиентов	SAGA	DESHL, SAGA
Архитектура	OGSA	OGSA
Протоколы передачи данных	SOAP	SOAP
Безопасность	X.509 (SSL)	X.509 (SSL)

Из таблицы видно, что существующие платформы для создания грид-приложений интеллектуального анализа данных имеют достаточно близкие технические характеристики. Так как UNICORE использует более свободную лицензию, а также в составе этой платформы уже присутствует графический клиент, было решено использовать для дальнейшей разработки платформу UNICORE.

При разработке программных систем обработки текстов на естественном языке необходимо решить

ряд проблем. Во-первых, в таких системах необходимо обеспечить механизмы взаимодействия между большим количеством разнородных компонентов, выполняющих разные виды анализа. Во-вторых, для таких систем требуются средства работы с большим количеством языковых ресурсов — размеченных корпусов, тезаурусов, моделей машинного обучения. В-третьих, поскольку алгоритмы обработки текстов обычно имеют высокую вычислительную сложность, для масштабируемости подобных систем, необходимо обеспечить возможность их распределения на несколько вычислительных узлов. В современных системах решение этих проблем берут на себя специализированные платформы анализа неструктурированной информации. Наиболее распространёнными на сегодняшний день являются платформы GATE и Apache UIMA. В табл. 3 представлены их сравнительные характеристики.

Таблица 3

Сравнительные характеристики платформ для обработки текстов на естественном языке

Название платформы	GATE	Apache UIMA
Лицензия	GPL, LGPL	Apache
Поддерживаемые языки программирования	Java	Java, C++
Поддержка распределенных вычислений	Нет	Да
Графический интерфейс для разработки приложения	Да	Да (плагин Eclipse)
Графический интерфейс для работы с разметкой	Да	Да (плагин Eclipse)
Средства для тестирования качества	Да	Да
Средства для профилирования скорости	Нет	Да

Большим преимуществом платформы UIMA перед GATE является возможность масштабирования за счет использования распределенных вычислений и наличие средств для профилирования анализаторов. Помимо этого, UIMA позволяет легко интегрировать в систему программные модули, реализованные как на языке Java, так и на языке C++. Это, с одной стороны, повышает гибкость разработки, за счет возможности внедрения в систему уже существующих компонентов на обоих языках, а с другой стороны, позволяет создавать эффективные модули за счет природы языка C++, ориентированной на высокопроизводительные вычисления. Исходя из этих преимуществ для разработки подсистемы анализа клинических текстов была выбрана платформа Apache UIMA.

Заключение

В работе предложена архитектура системы комплексного интеллектуального анализа медицинских данных, которая позволяет взаимодействовать с большим количеством разнородных источников структурированных и неструктурированных данных, включая клинические тексты на естественном языке, и выполнять анализ больших медицинских данных различными методами.

Была предложена программная платформа для создания системы комплексного интеллектуального анализа медицинских данных, а также выбрана платформа для создания подсистемы анализа медицинских текстов на русском языке, поддерживающая масштабирование и расширение при помощи сторонних модулей.

Предложенные архитектура и платформы лягут в основу разрабатываемой системы комплексного интеллектуального анализа медицинских данных в многопрофильном педиатрическом центре.

Литература

1. 2014 AA UMLS MeSH Russian source information. 2014 (окт.). http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSHRUS/index.html.
2. Международная классификация болезней 10-го пересмотра (МКБ-10). 2014 (окт.). http://mkb-10.com/.
3. *Agrawal R.*, *Imielinski T.*, *Swami A.* Mining association rules between sets of items in large databases // ACM SIGMOD Record / ACM. 1993. V. 22. P. 207–216.
4. American Psychiatric Association. The Diagnostic and Statistical Manual of Mental Disorders: DSM 5. Arlington, VA : American Psychiatric Association, 2013.
5. Apache OpenNLP. 2014 (окт.). https://opennlp.apache.org/index.html.
6. *Aronson A. R.* Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program // Proceedings of the AMIA Symposium / American Medical Informatics Association. 2001. P. 17–21.
7. Gene ontology: tool for the unification of biology / Michael Ashburner, Catherine A Ball, Judith A Blake et al. // Nature genetics. 2000. V. 25. № 1. P. 25–29.
8. Big data in health care: using analytics to identify and manage high-risk and high-cost patients / David W. Bates, Suchi Saria, Lucila Ohno-Machado et al. // Health Affairs. 2014. V. 33. № 7. P. 1123–1131.
9. A data mining system for infection control surveillance / S. E. Brossette, A. P. Sprague, W. T. Jones, S. A. Moser // Methods of information in medicine. 2000. V. 39, № 4/5. P. 303–310.
10. Data analysis services in the knowledge grid / Eugenio Cesario, Antonio Congiusta, Domenico Talia, Paolo Trunfio // Data Mining Techniques in Grid Computing Environments. 2008. P. 17–36.

11. Data mining approach to policy analysis in a health insurance domain / Young Moon Chae, Seung Hee Ho, Kyoung Won Cho et al. // International journal of medical informatics. 2001. V. 62. № 2. P. 103–111.
12. *Chen T.-J.*, *Chou L.-F.*, *Hwang S.-J.* Application of a data-mining technique to analyze coprescription patterns for antacids in Taiwan // Clinical therapeutics. 2003. V. 25. № 9. P. 2453–2463.
13. *Christensen L. M.*, *Haug P. J.*, *Fiszman M.* MPLUS: a probabilistic medical language understanding system // Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain. V. 3 / Association for Computational Linguistics. 2002. P. 29–36.
14. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model / Anni Coden, Guergana Savova, Igor Sominsky et al. // Journal of biomedical informatics. 2009. V. 42. № 5. P. 937–949.
15. *Cunningham H.* GATE, a general architecture for text engineering // Computers and the Humanities. 2002. V. 36. № 2. P. 223–254.
16. *Delen D.*, *Walker G.*, *Kadam A.* Predicting breast cancer survivability: a comparison of three data mining methods // Artificial intelligence in medicine. 2005. V. 34. № 2. P. 113–127.
17. *Ferrucci D.*, *Lally A.* UIMA: an architectural approach to unstructured information processing in the corporate research environment // Natural Language Engineering. 2004. V. 10. № 3–4. P. 327–348.
18. *Foster I.*, *Kesselman C.* The Grid 2: Blueprint for a new computing infrastructure. Elsevier, 2003.
19. *Foster I.*, *Maguire T.*, *Snelling D.* Ogsa wsrf basic profile 1.0. 2014. http://www.ogf.org/documents/GFD.72.pdf.
20. *Fox G. C.*, *Furmanski W.* PETAOPS and EXAOPS: Supercomputing on the web // Internet Computing, IEEE. 1997. V. 1. № 2. P. 38–46.
21. *Friedman C.* A broad-coverage natural language processing system // Proceedings of the AMIA Symposium / American Medical Informatics Association. 2000. P. 270–274.
22. Natural language processing in an operational clinical information system / Carol Friedman, George Hripcsak, William DuMouchel et al. // Natural Language Engineering. 1995. V. 1. № 01. P. 83–108.
23. A novel data mining system points out hidden relationships between immunological markers in multiple sclerosis / Maira Gironi, Marina Saresella, Marco Rovaris et al. // Immun Ageing. 2013. V. 10. № 1. http://www.biomedcentral.com/content/pdf/1742–4933–10–1.pdf.
24. Globus toolkit. 2014 (окт.). http://toolkit.globus.org/toolkit/.
25. *Todd R. Golub*, *Donna K. Slonim*, *Pablo Tamayo et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring // Science. 1999. V. 286. № 5439. P. 531–537.
26. *Harper P. R.* A review and comparison of classification algorithms for medical decision making // Health Policy. 2005. V. 71. № 3. P. 315–331.
27. A comparative study of classification methods for microarray data analysis / Hong Hu, Jiuyong Li, Ashley Plank et al. // Proceedings of the fifth Australasian conference on data mining and analystics / Australian Computer Society, Inc. 2006. V. 61. P. 33–37.
28. *Stanley M. Huff, Roberto A. Rocha, Clement J. McDonald et al.* Development of the logical observation identifier names and codes (LOINC) vocabulary // Journal of the American Medical Informatics Association. 1998. V. 5. № 3. P. 276–292.
29. ICD-10 Version:2010. 2014 (окт.). http://apps.who.int/classifications/icd10/browse/2010/en.
30. Kent ridge bio-medical dataset. 2014 (окт.). http://datam.i2r.a-star.edu.sg/datasets/krbd/.
31. *Lindberg D. A.*, *Humphreys B. L.*, *McCray A. T.* The unified medical language system // Methods of information in medicine. 1993. V. 32. № 4. P. 281–291.
32. Text analytics for life science using the unstructured information management architecture / R. Mack, Sougata Mukherjea, Aya Soffer et al. // IBM Systems Journal. 2004. V. 43. № 3. P. 490–515.
33. Mayo clinic. 2014. http://www.mayoclinic.org/.
34. Medical subject headings. 2014 (окт.). http://www.nlm.nih.gov/mesh/.
35. *Danielle L. Mowery, B. South, L. Christensen et al.* Task 2: ShARe/CLEF eHealth evaluation lab 2014 // CLEF 2014 Evaluation Labs and Workshop: Online Working Notes. 2014.
36. *Obenshain M. K.* Application of data mining techniques to healthcare data // Infection Control and Hospital Epidemiology. 2004. V. 25. № 8. P. 690–695.
37. Relational-situational method for intelligent search and analysis of scientific publications / Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Artem Shelmanov // Proceedings of the Workshop on Integrating IR technologies for Professional Search, in conjunction with the 35th European Conference on Information Retrieval (ECIR'13). V. 968. Moscow, Russia: CEUR Workshop Proceedings, 2013.
38. *Potter R.* Comparison of classification algorithms applied to breast cancer diagnosis and prognosis // Advances in Data Mining. 7th Industrial Conference, ICDM 2007, Leipzig, Germany, July 2007, Poster and Workshop Proceedings. 2007. P. 40–49.
39. Task 1: ShARe/CLEF eHealth evaluation lab 2013 / Sameer Pradhan, Noemie Elhadad, B South et al. // Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop. 2013.
40. Data mining meets grid computing: Time to dance? / Alberto Sánchez, Jesús Montes, Werner Dubitzky et al. // Data Mining Techniques in Grid Computing Environments. 2008. P. 1–16.
41. A data mining system for providing analytical information on brain tumors to public health decision makers / R. S. Santos, S. M. F. Malheiros, S. Cavalheiro, J. M. De Oliveira // Computer methods and programs in biomedicine. 2013. V. 109. № 3. P. 269–282.
42. *Guergana K. Savova, James J. Masanz, Philip V. Ogren et al.* Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications // Journal of the American Medical Informatics Association. 2010. V. 17. № 5. P. 507–513.
43. SemEval-2014 Task 7. 2014. http://alt.qcri.org/semeval2014/task7/index.php?id=task-description.
44. SemEval-2015 Task 14. 2014. http://alt.qcri.org/semeval2015/task14/index.php?id=task-description.

45. *Shah S.*, *Kusiak A.*, *Dixon B.* Data mining in predicting survival of kidney dialysis patients // Biomedical Optics 2003 / International Society for Optics and Photonics. 2003. P. 73–79.
46. SNOMED Clinical Terms. 2014. http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html.
47. SOAP Specifications — World Wide Web Consortium. 2014 (окт.). http://www.w3.org/TR/soap/.
48. *Sun W.*, *Rumshisky A.*, *Uzuner O.* Evaluating temporal relations in clinical text: 2012 i2b2 challenge // Journal of the American Medical Informatics Association. 2013. V. 20. № 5. P. 806–813.
49. *Tassy O.*, *Pourquié O. Manteia*, a predictive data mining system for vertebrate genes and its applications to human genetic diseases // Nucleic acids research. 2014. V. 42. № D1. P. D882–D891.
50. UNICORE — Distributed computing and data resources. 2014. http://www.unicore.eu/.
51. *Ozlem Uzuner, Brett R. South, Shuying Shen, Scott L. DuVall.* 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text // Journal of the American Medical Informatics Association. 2011. P. 552–556.
52. Gene expression profiling predicts clinical outcome of breast cancer / Laura J van't Veer, Hongyue Dai, Marc J Van De Vijver et al. // Nature. 2002. V. 415. № 6871. P. 530–536.
53. Weka 3: Data mining software in java. 2014. http://www.cs.waikato.ac.nz/ml/weka/.
54. MedEx: a medication information extraction system for clinical narratives / Hua Xu, Shane P Stenner, Son Doan et al. // Journal of the American Medical Informatics Association. 2010. V. 17. № 1. P. 19–24.
55. Data mining in healthcare and biomedicine: a survey of the literature / Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov et al. // Journal of medical systems. 2012. V. 36. № 4. P. 2431–2448.
56. Piloting IBM Watson Oncology within Memorial Sloan Kettering's regional network. / Marjorie Glass Zauderer, Ayca Gucalp, Andrew S Epstein et al. // ASCO Annual Meeting Proceedings. 2014. V. 32. P. e17653.
57. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system / Qing T Zeng, Sergey Goryachev, Scott Weiss et al. // BMC medical informatics and decision making. 2006. V. 6. № 30.

Баранов Александр Александрович. Директор ФГБНУ НЦЗД. Д. мед. н., профессор, академик РАН. Окончил в 1964 г. Казанский государственный медицинский институт. Количество печатных работ: 555. Область научных интересов: педиатрия, общественное здоровье и здравоохранение. E-mail: baranov@nczd.ru

Намазова-Баранова Лейла Сеймуровна. Зам. директора ФГБНУ НЦЗД, директор НИИ ППиВЛ ФГБНУ НЦЗД. Д. мед. н., профессор, член-корр. РАН. Окончила в 1987 г. 2-ой МОЛГМИ им. Н. И. Пирогова в 1987 г. Количество печатных работ: 598. Область научных интересов: педиатрия, общественное здоровье и здравоохранение, аллергология-иммунология. E-mail: namazova@nczd.ru

Смирнов Иван Валентинович. С. н. с. ИСА РАН. К. ф.-м. н. Окончил в 2003 г. РУДН. Количество печатных работ: 43. Область научных интересов: анализ естественного языка, интеллектуальный анализ данных, интеллектуальные поисковые машины. E-mail: ivs@isa.ru

Девяткин Дмитрий Алексеевич. М. н. с. ИСА РАН. Окончил в 2011 г. Рыбинскую государственную авиационную технологическую академию им. П. А. Соловьёва. Количество печатных работ: 7. Область научных интересов: методы интеллектуального анализа данных, анализ больших данных, анализ тональности текстов. E-mail: devyatkin@isa.ru

Шелманов Артем Олегович. М. н. с. ИСА РАН. Окончил в 2011 г. МИФИ. Количество печатных работ: 7. Область научных интересов: искусственный интеллект, компьютерная лингвистика, информационно-аналитические системы, машинное обучение. E-mail: shelmanov@isa.ru

Вишнева Елена Александровна. Зав. отделом стандартизации и клинической фармакологии ФГБНУ НЦЗД. К. мед. н. Окончила в 2004 г. ММА им. И. М. Сеченова. Количество печатных работ: 87. Область научных интересов: педиатрия, общественное здоровье и здравоохранение, аллергология-иммунология. E-mail: vishneva@nczd.ru

Антонова Елена Вадимовна. Зав. отделом ФГБНУ НЦЗД. Д. мед. н. Окончила в 1987 г. 2-ой МОЛГМИ им. Н. И. Пирогова. Количество печатных работ: 93. Область научных интересов: педиатрия, общественное здоровье и здравоохранение. E-mail: antonova@nczd.ru

Смирнов Владимир Иванович. Зам. директора НИИ ППиВЛ ФГБНУ НЦЗД. К. э. н. Окончил в 1996 г. Государственную академию управления им. С. Орджоникидзе. Количество печатных работ: 11. Область научных интересов: информационные технологии, организация здравоохранения. E-mail: support@nczd.ru

Латышев Андрей Валерьевич. С. н. с. ИСА РАН. Окончил в 1985 г. МАИ. К. т. н. Количество печатных работ: 10. Область научных интересов: информационно-аналитические системы различного применения, системотехника, информационные, сетевые и телекоммуникационные технологии. E-mail: andrey.latyshev@gmail.com