УТВЕРЖДАЮ

Генеральный директор
АО «Авикомп Сервисез»

_____ В.П. Клинцов

« » _____ 2021 г.

СИСТЕМА ПОДДЕРЖКИ ПРИНЯТИЯ ВРАЧЕБНЫХ РЕШЕНИЙ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ КОНТРОЛЯ (ОЦЕНКИ) КАЧЕСТВА ЛЕЧЕНИЯ В СООТВЕТСТВИИ С УТВЕРЖДЕННЫМИ КЛИНИЧЕСКИМИ РЕКОМЕНДАЦИЯМИ (СППВР ИИ)

Инв. № подл.	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата

Описание применения

ЛИСТ УТВЕРЖДЕНИЯ

АВКС.00223-01 31 01-ЛУ

Руководитель проекта

А.Ю. Лысенко

«___»_____2021 г.

Нормоконтролер

А.Н.Чепак

«___»_____2021 г.

2021

Утвержден
АВКС.00223-01 31 01-ЛУ

СИСТЕМА ПОДДЕРЖКИ ПРИНЯТИЯ ВРАЧЕБНЫХ РЕШЕНИЙ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ КОНТРОЛЯ (ОЦЕНКИ) КАЧЕСТВА ЛЕЧЕНИЯ В СООТВЕТСТВИИ С УТВЕРЖДЕННЫМИ КЛИНИЧЕСКИМИ РЕКОМЕНДАЦИЯМИ (СППВР ИИ)

Описание применения

АВКС.00223-01 31 01

Листов 52

Инв. № подл.	Подп. и дата	Инв. № дубл.	Взам. инв. №	Подп. и дата

2021

АННОТАЦИЯ

Настоящий документ описывает назначение, условия применения, основные характеристики, задачи и методы их решения, массив данных специального программного обеспечения «Система поддержки принятия врачебных решений с использованием технологий искусственного интеллекта для контроля (оценки) качества лечения в соответствии с утвержденными клиническими рекомендациями (СППВР ИИ, далее – Система).

Основание для разработки Системы: Договор на выполнение научно-исследовательской и опытно-конструкторской работы от 18.12.2019 г. № 80ГРЦТС10-D5/56141, заключенный между «Фондом содействия развитию малых форм предприятий в научно-технической сфере» (Фонд содействия инновациям) и АО «Авикомп Сервисез».

Организация-разработчик: АО «Авикомп Сервисез».

При разработке настоящего документа использовались ГОСТ 19.105-78 «Единая система программной документации. Общие требования к программным документам», ГОСТ 19.106-78 «Единая система программной документации. Требования к программным документам, выполненным печатным способом», ГОСТ 19.502-78 «Единая система программной документации. Описание применения. Требования к содержанию и оформлению».

СОДЕРЖАНИЕ

1 НАЗНАЧЕНИЕ ПРОГРАММЫ

Программа является информационно-справочной системой поддержки принятия врачебных решений путём выявления релевантной информации, находящейся в гетерогенной базе данных (ГБД). Видами релевантной информации, извлекаемой на основании ввода исходных данных о состоянии пациента, являются:

– вероятные диагнозы;

– рекомендованные методы диагностики;

– рекомендованное лечение.

Исходными данными являются:

– симптомы;

– синдромы;

– диагнозы всех типов (ранее поставленные и подтверждённые, хронические, предполагаемые в данный момент);

– методы диагностики, проведённая диагностика, результаты проведённой диагностики;

– методы лечения, применяемые в данный момент, планируемые.

Программа предназначена для применения при оказании первичной врачебной помощи, специализированной врачебной помощи, высокотехнологичной врачебной помощи, паллиативной медицинской помощи в условиях стационара или амбулатории.

Ретроспективное использование системы может выявить врачебные ошибки:

– излишние шаги, неверно назначенные диагностические и лечебные процедуры,

– неверно поставленные диагнозы,

– неправильные методы диагностики и лечения,

– неверно назначенные лекарства.

1.1 Возможности программы

Основными возможностями Системы являются:

– сбор данных справочников и исходных документов – формирование информационной базы;

– извлечение метаданных, в том числе объектов и связей путём семантической обработки справочников и исходных документов;

– формирование ГБД;

– анализ клинических данных пациента, извлечение релевантных данных из ГБД и предоставление их пользователю.

1.2 Основные характеристики программы

1.2.1 Состав Системы

Система состоит из следующих компонентов (рис. 1):

– подсистема сбора исходных документов (ПС Сбора);

– подсистема извлечения метаданных (ПС Извлечения);

– подсистема ввода данных (ПС Ввода);

– подсистема ведения нормативно-справочной информации (ПС Ведения НСИ);

– подсистема обучения (ПС Обучения);

– подсистема хранения данных (ПС Хранения);

– подсистема предоставления аналитических данных (ПС Предоставления АД);

– подсистема настройки и формирования отчетных форм (ПС ФОФ);

– подсистема обратной связи (ПС ОС);

– подсистема администрирования (ПС Администрирования, не показана на рис. 1).

В систему также входят два дополнительных специализированных модуля:

– модуль ускорения вычислений «DataMonitor» – интерфейс настройки визуализаций для администратора системы, включающий программное средство «Монитор данных» («DataMonitor»).

– модуль «Семантический индекс для PostgreSQL» – модуль семантической обработки текстов.

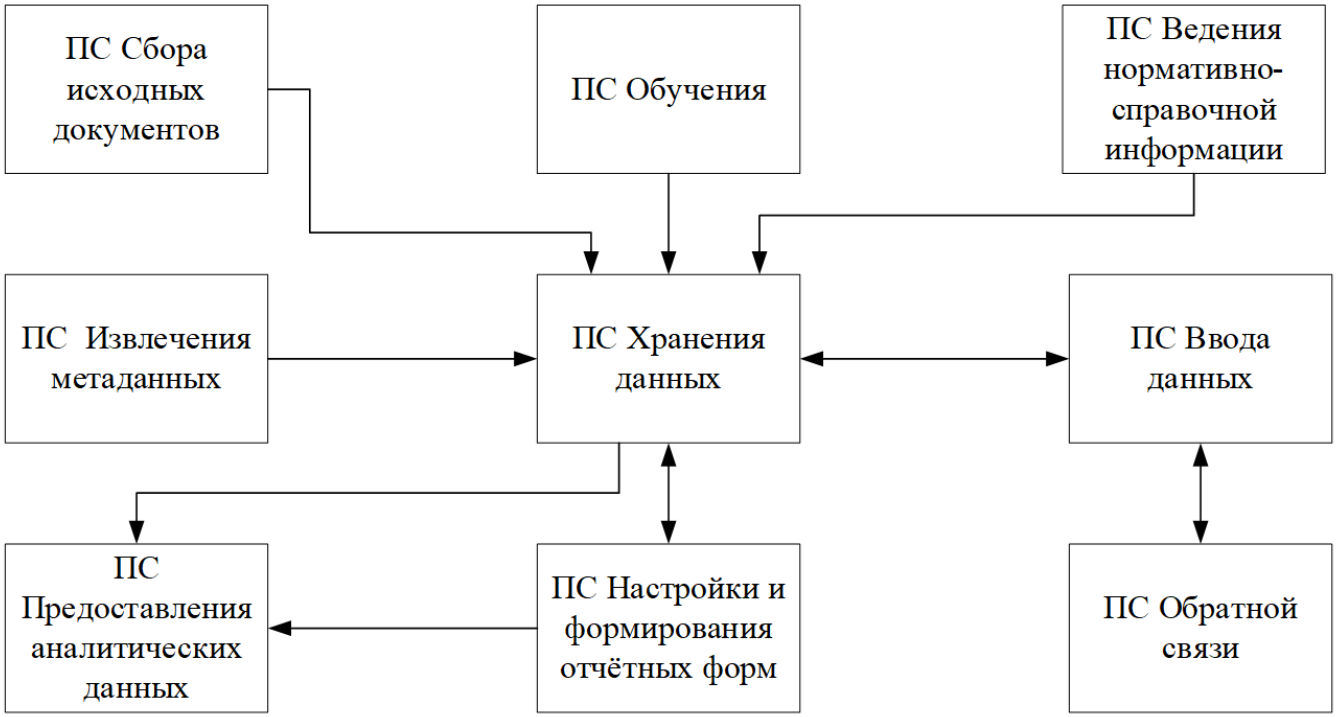

Рисунок 1 - Структура и взаимодействие Системы

1.2.2 Подсистема сбора исходных документов

ПС Сбора представляет собой набор скриптов, которые осуществляют загрузку данных в ПС Хранения из различных источников с целью их дальнейшей семантической обработки.

Могут загружаться как структурированные, так и не структурированные данные в различных форматах. Источником может быть:

– набор файлов в папке файловой системы,

– таблица или набор таблиц реляционной базы данных,

– отдельный структурированный csv-файл, содержащий текстовые и цифровые данные.

При всём разнообразии источников как по форматам представления, так и по объёмам, целевым способом и местом хранения данных ПС Сбора являются таблицы реляционной базы данных, где подлежащие семантической обработке

текстовые данные размещаются в определённом формате, поддерживаемом модулем АВКС.00309-01 «Семантический индекс для PostgreSQL».

Скрипт-обработчик считывает данные из таблицы реляционной базы данных, в которой в унифицированном виде хранятся исходные тексты для обработки, и записывает метаданные, полученные в результате обработки, в результирующие таблицы реляционной базы данных.

1.2.3 Подсистема хранения данных

ПС Хранения состоит из нескольких программных систем, обеспечивающих хранение и обработку данных. Помимо реляционной базы данных PostgreSQL (https://www.postgresql.org/), в ПС Хранения входит ряд программных и технических компонент, обеспечивающих ускорение обработки таблиц большого размера, а также программная система, поддерживающая нереляционную модель данных, в которой объекты и связи имеют равное значение, и предназначенная для поддержки больших наборов данных.

1.2.4 Подсистема ввода данных

ПС Ввода состоит из следующих компонент:

– серверного компонента гетерогенной базы знаний (ГБД);

– таблиц БД PostgreSQL с данными по взаимодействию лекарственных средств;

– серверного компонента, написанного на языке скриптов PHP, для выполнения и обработки клиентских запросов;

– веб-клиента;

– клиентского API для получения данных из БД;

– клиентской системы выполнения запросов для обработки и получения данных из ГБД, реализованной с помощью протокола обмена данных websocket.

Пользователь, выполняющий действия в веб-интерфейсе (ввод, выбор данных), инициализирует протокол обмена информацией с ГБД, реализуемый веб-

клиентом. Таким образом, с помощью инструментов веб-интерфейса пользователь может формировать сложные запросы, и получать информацию по заданным пользователем критериям.

Так как веб-клиент взаимодействует с серверными компонентами посредством API или по протоколу websocket, часть обработки данных (группировка, сортировка) производится на стороне клиента.

Перечень сценариев, инициируемых пользователем через веб-клиент:

1) Выполнить ввод симптомов или показаний.

2) Получить список симптомов, часто встречающихся с отобранными ранее.

3) Получить группированный список предполагаемых диагнозов.

4) Получить дополнительную информацию по диагнозу:

– похожие диагнозы,

– диагностические признаки,

– медикаменты для лечения,

– встречающиеся симптом.

1.2.5 Подсистема ведения нормативно-справочной информации

ПС Ведения НСИ представляет собой набор скриптов и программ, предназначенных для формирования специальных наборов данных для загрузки в ГБД.

Для формирования могут быть использованы данные в таблицах реляционной БД, а также дистрибутивные файлы справочных и иных систем. При этом данные из дистрибутивных файлов справочных и других систем могут сначала загружаться в реляционную базу данных и потребляться ПС Ведения НСИ из таблиц реляционной системы. Такие таблицы в реляционной БД формируются подсистемой сбора исходных документов (ПС Сбора) в результате обработки исходных дистрибутивов медицинских библиотек, загруженных и обработанных текстовых материалов, из которых извлечена метаинформация.

Указанные скрипты и программы формируют загрузочные файлы в специальном формате загрузчика ГБД с учетом применяемой схемы ГБД, а именно: применяемых объектов, связей, свойств объектов и связей, меток объектов.

Комплект скриптов может поставляться в специальном архиве: sppvr_ii_nsi_scripts.gz.tg .

1.2.6 Подсистема извлечения метаданных

ПС Извлечения состоит из следующих компонент:

– серверного компонента, написанного на языке скриптов PHP, для выполнения и обработки клиентских запросов;

– серверного компонента ГБД;

– таблиц БД PostgresQL с данными по взаимодействию лекарственных средств;

– веб-клиента;

– клиентского API для взаимодействия с БД;

– клиентской системы поддержки запросов для получения данных из ГБД, реализованной с помощью протокола обмена данных websocket;

Пользователь, выполняющий действия в веб-интерфейсе (ввод, фильтрация, выбор данных), инициализирует клиентское API либо протокол обмена информацией с ГБД, реализуемые веб-клиентом. Серверные компоненты получают данные от клиента, формируют SQL-запросы, извлекая с их помощью необходимые данные из БД. В случае с серверным компонентом для работы с ГБД, запросы формируются непосредственно на стороне клиента и отправляются по протоколу websocket. Таким образом, с помощью инструментов веб-интерфейса пользователь может формировать сложные запросы и получать информацию по заданным пользователем критериям.

Так как веб-клиент взаимодействует с серверными компонентами посредством API или по протоколу websocket, часть обработки данных (группировка, сортировка) производится на стороне клиента.

Перечень сценариев, инициируемых пользователем через веб-клиент:

1) Подготовить запрос для поиска документов по определенным категориям (подкатегориям).

2) Получить список кластеров документов, сгруппированных общими понятиями с использованием интеллектуальной системы кластеризации документов.

3) Получить документ с размеченными медицинскими понятиями.

4) Подготовить произвольный документ и применить к нему метод разметки медицинскими понятиями.

5) Получить дополнительную информацию по размеченному элементу в документе.

1.2.7 Подсистема настройки и формирования отчетных форм – Модуль ускорения вычислений «DataMonitor»

ПС ФОФ предназначена для формирования интерактивных отчетных форм, в качестве которых использованы аналитические панели «Монитора данных». «Монитор данных» является частью «Модуля ускорения вычислений «DataMonitor». который состоит из следующих модулей:

– модуль «Монитор данных»;

– модуль ускорителя GPU для Postgres SQL.

Информация о данном программном модуле представлена в документации АВКС.00305-01 «Модуль ускорения вычислений «DataMonitor».

1.2.8 Подсистема обучения

ПС Обучения предназначена для формирования обучающих наборов данных и тренировки нейронной сети NLP-модуля:

– создание датасета (корпуса текстов) для тренировки майнеров – нейронных сетей;

– тренировка унимодальных майнеров, использующих обученную нейронную сеть, для каждого отдельного типа объектов: «лекарственные препараты», «состояния/заболевания», «синдромы», «анатомия» и т.п.

С помощью программы пользователь может:

– создавать словники (словари-справочники);

– создавать корпуса текстов (наборы текстовых документов);

– запускать обучение майнера нейронной сети;

– выполнять оценку качества работы обученного майнера.

Обученные майнеры используются подсистемой извлечения метаданных (ПС Извлечения).

1.2.9 Подсистема обратной связи

Функциональность ПС ОС реализована в составе подсистемы ввода данных (ПС Ввода). Эта функциональность предусматривает в случае, когда в процессе диалога с пользователем подсистемы что-то пошло не так, и у пользователя возникли вопросы к предоставляемой Системой информации, следующие действия пользователя:

– зафиксировать состояние интерфейса путём создания скриншота окна с диалогом;

– написать сопровождающее сообщение;

– отправить сообщение вместе с зафиксированным состоянием интерфейса группе поддержки на специальный адрес.

Группа поддержки по результатам анализа полученного сообщения выполняет корректирующие мероприятия.

1.3 Алгоритм программы

Система решает задачи, реализуемые её программными компонентами:

– подсистемой сбора исходных документов (ПС Сбора);

– подсистемой извлечения метаданных (ПС Извлечения);

– подсистемой ввода данных (ПС Ввода);

– подсистемой ведения нормативно-справочной информации (ПС Ведения НСИ);

– подсистемой обучения (ПС Обучения);

– подсистемой хранения данных (ПС Хранения);

– подсистемой предоставления аналитических данных (ПС Предоставления АД);

– подсистемой настройки и формирования отчетных форм (ПС ФОФ);

– подсистемой обратной связи (ПС ОС);

– подсистемой администрирования (ПС Администрирования).

В настоящем разделе описаны алгоритмы, реализованные перечисленными выше подсистемами.

1.3.1 Описание алгоритма работы подсистемы сбора исходных документов

ПС Сбора предназначена для преобразования и записи загружаемых данных в унифицированное представление, пригодное для дальнейшей обработки, выполняемой программным модулем АВКС.00309-01 «Семантический индекс для PostgreSQL».

Унифицированным представлением является таблица реляционной базы данных, где подлежащие семантической обработке текстовые данные размещаются в определённом формате, поддерживаемом модулем «Семантический индекс для PostgreSQL».

Скрипт-обработчик считывает данные из таблицы реляционной базы данных, в которой в унифицированном виде хранятся исходные тексты для обработки, и записывает метаданные, полученные в результате обработки, в результирующие таблицы реляционной базы данных.

Более подробная информация о функционировании модуля «Семантический индекс для PostgreSQL» представлена в документации на модуль.

1.3.2 Описание алгоритма работы подсистемы извлечения метаданных

К моменту начала использования ПС Извлечения основная часть метаданных уже извлечена – подсистемой сбора исходных документов сформирован

семантический индекс. В подсистеме извлечения метаданных реализовано обращение к функциям извлечения сущностей, которые выполняются в случае разметки стороннего документа, не размещённого в ПС Хранения. Основная часть функциональности ПС Извлечения – отображение размеченных документов для пользователя-аналитика, включая различные типы поиска информации в документах.

Обобщённый алгоритм работы ПС Извлечения может быть представлен как последовательное выполнение следующих действий:

1) При инициализации URL-адреса страницы ПС Извлечения в браузере пользователя Система формирует HTML-файл со страницей графического интерфейса, для отображения которого производится формирование и загрузка ресурсов из серверных компонентов. Графический интерфейс формируется с помощью скриптов и правил, примененных для CSS, JS и HTML5.

2) В процессе взаимодействия пользователя и графического интерфейса ПС Извлечения осуществляются функции поиска, фильтрации и выбора данных. По результатам выполнения этих действий пользователя формируются запросы к соответствующим серверным компонентам ПС Извлечения. Результаты ответа на них выводятся на графический интерфейс в соответствии с настроенными шаблонами модулей пользовательского интерфейса.

1.3.3 Описание алгоритма работы подсистемы ввода данных

Обобщённый алгоритм работы ПС Ввода может быть представлен как последовательное выполнение следующих действий:

1) При инициализации URL-адреса страницы ПС Ввода в браузере пользователя система формирует HTML-файл со страницей графического интерфейса, формируемого с помощью скриптов и правил, примененных для CSS, JS и HTML5.

2) В процессе взаимодействия пользователя и графического интерфейса ПС Ввода осуществляются функции поиска, фильтрации, выбора данных и их удаление. Во время выполнения этих действий пользователя формируются запросы

к соответствующим серверным компонентам ПС Хранения. Ответ выводится на графический интерфейс в соответствии с настроенными шаблонами модулей пользовательского интерфейса.

1.3.4 Описание алгоритма работы подсистемы ведения нормативно-справочной информации

ПС Ведения НСИ представляет собой набор скриптов и программ, предназначенных для формирования специальных наборов данных для загрузки в ГБД.

Для формирования могут быть использованы данные в таблицах реляционной БД, а также дистрибутивные файлы справочных и иных систем. При этом данные из дистрибутивных файлов справочных и других систем могут сначала загружаться в реляционную базу данных и потребляться ПС Ведения НСИ из таблиц реляционной системы. Такие таблицы в реляционной БД формируются подсистемой сбора исходных документов (ПС Сбора) в результате обработки исходных дистрибутивов медицинских библиотек, загруженных и обработанных текстовых материалов, из которых извлечена метаинформация.

Указанные скрипты и программы формируют загрузочные файлы в специальном формате загрузчика ГБД с учетом применяемой схемы ГБД, а именно: применяемых объектов, связей, свойств объектов и связей, меток объектов.

Для каждого справочника используется специальный скрипт, формирующий загрузочный файл для данных этого справочника.

После того, как загрузочные файлы подготовлены, специальная программа-загрузчик формирует базу знаний.

1.3.5 Описание алгоритма работы подсистемы обучения

ПС Обучения автоматизирует следующую последовательность действий:

– формирование (расширение) усечённого набора данных;

– формирование (расширение) лингвистического графа;

– формирование дополненного словаря;

– разбиение словаря;

– формирование обучающего набора данных;

– формирование тестового набора данных;

– устранение утечки данных;

– обучение майнера;

– контроль майнера на полноту и точность извлечения именованных сущностей;

– расширение обучающего набора данных.

1.3.5.1 Формирование (расширение) усечённого набора данных

Входом для операции служит большой набор данных (НД). На основании словаря, содержащего некоторое множество терминов, характеризующих предметную область, Система отбирает из большого набора данных документы, содержащие термины из словаря, и помещает отобранные документы в выходной усеченный набор данных. В результате выполнения последующих итераций усечённый набор данных будет расширяться за счет расширения словаря. Итеративное расширение словаря предполагается за счет ручного отбора аналитиком терминов из разметки, выполненной обученным майнером.

1.3.5.2 Формирование (расширение) лингвистического графа

Семантический граф представляет собой взвешенный граф (каждому ребру которого поставлено в соответствие некое значение – вес ребра), вершинами которого являются корпуса документов. Наличие ребра между двумя вершинами означает тот факт, что термины семантически связаны между собой, а вес ребра является численным значением семантической близости двух терминов, которые соединяет данное ребро. Формирование (расширение) лингвистического графа может выполняться либо для полного, либо для усечённого набора данных в

зависимости от размера этих наборов и имеющейся в наличии вычислительной мощности.

1.3.5.3 Формирование дополненного словаря – модуль формирования тематических словарей

Аналитик дополняет словарь, анализируя семантический граф, представляющий семантически близкие (то есть, упоминаемые в близких контекстах) термины.

1.3.5.4 Разбиение словаря

Поскольку целью алгоритма является обучение майнера, необходимо создать два набора данных:

– обучающий набор данных, содержащий пример выделения терминов (пример разметки текстов терминами соответствующих типов), выделению которых должен обучиться майнер;

– тестовый набор данных, также содержащий выделение (разметку), по которому будет проверено качество работы (полнота и точность выделения терминов заданного типа) уже обученного майнера.

Соответственно этому словарь должен быть случайным образом разбит на две части:

– обучающую – термины, которые будут использованы для обучения (примерно 80% от общего количества терминов словаря);

– тестовую – термины, которые будут использованы для контроля обучения (примерно 20% от общего количества терминов словаря).

1.3.5.5 Формирование обучающего набора данных

Для формирования обучающего набора данных ПС Обучения просматривает усечённый набор данных и помещает в набор данных для обучения те предложения

из документов усечённого набора данных, которые содержат термины, вошедшие в словарь образов для обучения майнера.

1.3.5.6 Формирование тестового набора данных

Для формирования набора данных для тестирования обучения майнера ПС Обучения просматривает усечённый набор данных и помещает в набор данных для тестирования те предложения из документов усечённого набора данных, которые содержат термины, вошедшие в словарь образов для тестирования майнера.

1.3.5.7 Устранение утечки данных

Утечкой данных (англ. «data leak») в машинном обучении называется дублирование тестовых данных в данных для обучения. В сформированном наборе предложений для обучения (обучающем наборе данных) не должно быть предложений, содержащих выделенные термины, вошедшие в словарь для тестирования (словарь образцов для тестирования), и наоборот – в сформированном наборе предложений для тестирования (тестовом наборе данных) не должно быть предложений, содержащих выделенные термины, вошедшие в словарь для обучения (словарь образцов для обучения).

При выполнении операции формирования обучающего и тестового наборов данных выполняется поиск терминов, принадлежащих соответствующей части словаря, а отсутствие терминов другой части словаря не контролируется. Поэтому всегда имеется вероятность того, что отобранные предложения могут содержать термины другой части словаря, например, в перечислениях различных объектов одинаковых типов, при сопоставлении различных объектов одного типа, а также в каких-то других случаях.

Для устранения утечки алгоритм предусматривает просмотр обучающего набора на наличие в предложениях терминов словаря для тестирования и удаление из набора данных предложений, где такие термины обнаружены. И наоборот, просмотр тестового набора данных на наличие в предложениях терминов словаря

для обучения и удаление из набора данных предложений, где словарные термины для обучения обнаружены.

1.3.5.8 Обучение майнера

Майнер представляет собой нейронную сеть. На вход алгоритма подаётся обучающий НД. Выходом является матрица коэффициентов, которая записывается в БД.

1.3.5.9 Контроль майнера на полноту и точность извлечения именованных сущностей

Для выявления параметров работы обученного майнера выполняется контроль на полноту и точность извлечения того типа именованных сущностей, для которого были сформированы обучающий и тестовый наборы данных и выполнено обучение (тренировка).

На вход обученному майнеру подаётся тестовый набор данных, майнер выполняет выделение сущностей в этом наборе; результаты выделения сущностей майнером сравниваются с эталонной разметкой – рассчитываются метрики полноты и точности выделения объектов.

Точность можно интерпретировать как долю объектов, определённых майнером как принадлежащих типу терминов, для которых был обучен майнер, и при этом действительно являющихся таковыми, а полнота показывает, какую долю терминов определяемого майнером типа из всех терминов данного типа нашел майнер.

Рассчитанные значения полноты и точности сравниваются с заданными. В случае, если данные метрики оказались меньше заданных пороговых значений, то есть, меньше, чем задано пороговым значением полноты, которое было выделено из имеющихся именованных сущностей в тестовом наборе данных, и среди выделенных присутствует больше именованных сущностей других типов, чем было задано пороговым значением точности, выполняется следующая итерация пополнения обучающего и тестового наборов данных. Если полнота и точность,

рассчитанные по результатам обработки тестового набора данных, удовлетворяют заданным пороговым значениям – майнер считается обученным и передаётся в ПС Извлечения для использования.

1.3.5.10 Расширение обучающего набора данных

Расширение обучающего НД может выполняться путём расширения словаря с целью вовлечения в обучающий НД большего количества предложений с образцами, в которых использовано большее количество именованных сущностей заданного типа, а также увеличения количества предложений без изменения состава словаря.

Система предусматривает оба варианта расширения состава обучающего НД:

– в первом случае пользователь-аналитик должен пополнить словарь терминами, используя для этого результаты разметки имеющимся майнером и сформировав новый обучающий (и тестовый, если необходимо) набор данных большего объёма;

– во втором случае пользователь может добавить новые предложения в усечённый набор данных, не изменяя объём словаря.

1.3.6 Описание алгоритма работы подсистемы хранения данных

ПС Хранения состоит из программных компонент, обеспечивающих хранение и обработку данных, в том числе PostgreSQL (https://www.postgresql.org/).

PostgreSQL является большой развитой системой, предназначенной для хранения и обработки больших объёмов данных. Алгоритмические аспекты, связанные с реализацией этого назначения, подробно описаны в её документации. Ссылка на источник документации приведена выше.

1.3.7 Описание алгоритма работы подсистемы предоставления аналитических данных)

ПС Предоставления АД осуществляется порталом. Страницы пользовательского портала формируются серверной частью на основе описаний, заранее созданных с пользователем-аналитиком и имеющихся в специальной схеме портала. Конфигурационные данные, содержащиеся в указанной схеме, описывают конфигурацию портала для определённой группы пользователей. В момент авторизации пользователя на портале происходит определение принадлежности пользователя к группе. Сквозная авторизация пользователя обеспечивается стандартными механизмами операционной системы. Для формирования структуры портала используется соответствующая группе конфигурация. Конфигурация портала описывает:

– структуру страниц;

– состав страниц;

– содержание страниц;

– запросы к БД для формирования списков, отображаемых на страницах;

– данные виджетов, внедрённых на страницы.

Обобщённый алгоритм работы пользовательского портала может быть представлен как последовательное выполнение следующих действий:

1) При инициализации основной страницы Система создает и формирует сессию пользователя, сохраняя в процессе работы в серверном окружении данные, необходимые для работы страниц разделов пользовательского портала в браузере.

2) Пользователь с помощью веб-браузера, указывая URL-портала, открывает страницу портала.

3) Система формирует HTML-файл со страницей графического интерфейса пользовательского портала, для загрузки которого производится формирование и загрузка ресурсов (шаблоны CSS, JS, HTML5), реализующих основной загрузчик, и шаблоны графического интерфейса страниц пользовательского портала.

4) В процессе взаимодействия пользователя и страниц пользовательского портала осуществляются функции поиска, фильтрации и выбора данных. По результатам выполнения этих действий пользователя формируются запросы к соответствующим компонентам Системы. Результаты ответа на них выводятся на страницы пользовательского портала в соответствии с настроенными шаблонами страниц разделов пользовательского портала.

Пользовательский портал позволяет описать конфигурации портала для различных групп пользователей.

Обобщённый алгоритм работы конструктора пользовательского портала может быть представлен как последовательное выполнение следующих действий:

1) Открытие интерфейса конструктора пользовательского портала в браузере пользователя.

2) Система формирует HTML-файл с главной страницей графического интерфейса конструктора пользовательского портала, при загрузке которого производится формирование и скачивание ресурсов (шаблоны CSS, JS, HTML5), реализующих основной загрузчик и шаблоны графического интерфейса страниц конструктора пользовательского портала.

3) При инициализации основной страницы Система создает и формирует сессию пользователя, сохраняя в процессе работы в серверном окружении данные, необходимые для работы страниц разделов конструктора пользовательского портала в браузере.

4) В процессе взаимодействия администратора и страниц конструктора пользовательского портала осуществляются функции настройки страниц разделов пользовательского портала, с возможностью поиска и фильтрации настроенных страниц и их элементов. По результатам выполнения этих действий администратора сохраняются настройки элементов страниц пользовательского портала, которые используются для формирования страниц пользовательского портала.

1.3.8 Описание алгоритма работы подсистемы настройки и формирования отчетных форм

Обобщённый алгоритм работы ПС ФОФ (монитора данных) может быть представлен как последовательное выполнение следующих действий:

1) Открытие интерфейса панели в браузере пользователя.

2) Система формирует HTML-файл с главной формой графического интерфейса монитора данных, при загрузке которого производится формирование и скачивание ресурсов (шаблоны CSS, JS, HTML5), реализующих основной загрузчик и виджеты графического интерфейса.

3) При инициализации основного виджета Система создает и формирует сессию пользователя, сохраняя в процессе работы в серверном окружении данные, необходимые для работы виджетов в браузере. В процессе взаимодействия загрузчик автоматически выполняет формирование дополнительных ресурсов и загрузку дополнительных виджетов.

4) Производится загрузка главной формы графического интерфейса монитора данных.

5) Система выполняет загрузку настроек рабочего пространства пользователя, после чего выполняет визуализацию дерева объектов рабочего пространства монитора данных, включающего: структуру папок, источники и объекты источников данных, настройки кубов и срезов, настройки визуализаций.

6) Система ожидает дальнейших действий пользователя:

6.1) Изменение структуры дерева рабочего пространства пользователя монитора данных.

6.2) Создание и редактирование источника данных.

6.3) Создание и редактирование куба с показателями.

6.4) Создание и редактирование визуализации.

7) Сохранение изменений производится автоматически, все дальнейшие действия выполняются в рамках шага 6.

1.3.8.1 Изменение структуры дерева рабочего пространства пользователя монитора данных

Обобщённый алгоритм работы модуля в части изменении структуры дерева рабочего пространства пользователя монитора данных может быть представлен как последовательное выполнение следующих действий:

1) Система ожидает от пользователя одно из следующих действий:

– создание папки или перемещение объекта в папку рабочего пространства пользователя;

– удаление папки или объекта рабочего пространства пользователя.

2) Согласно выбранному действию, Система модифицирует в памяти структуру рабочего пространства монитора данных.

3) Система выполняет автоматическое сохранение изменений файлов рабочего пространства пользователя на диск. На диске рабочее пространство сохранено в виде двух типов файлов:

– файл рабочего пространства в формате JSON;

– файл ресурса рабочего пространства пользователя в формате JSON (настройки куба, настройки визуализаций).

4) В соответствии с результатами выполнения выбранного пользователем действия Система обновляет виджеты графического интерфейса, перерисовывая изменения дерева рабочего пространства.

5) Возврат к шагу 1 или к пункту 1.3.8, шагу 6.

1.3.8.2 Создание и редактирование источника данных

Обобщённый алгоритм работы модуля в части создания и редактирования источника данных может быть представлен как последовательное выполнение следующих действий:

1) Пользователь выбирает действие «Создание и редактирование источника данных».

2) Система подготавливает и загружает в браузер виджеты редактирования источника данных в виде набора ресурсов (шаблоны JS, CSS, HTML5).

3) Если производится редактирование существующего источника, система инициализирует виджеты заданными параметрами источника, если нет – производится инициализация виджетов значениями по умолчанию.

4) Пользователь задает параметры источника (название, строка и параметры соединения с источником), после чего отправляет запрос на тестовое подключение и загрузку схемы.

5) Система создает в памяти объект рабочего пространства с настройками источника, выполняет автоматическое сохранение обновленного файла рабочего пространства на диске.

6) Система выполняет тестовое подключение к источнику, в случае ошибки подключения формируется и выбрасывается исключение с последующей визуализацией в браузере.

7) Система выполняет загрузку схемы источника (перечень и параметры таблиц, визуализаций и столбцов), в случае ошибки подключения формируется и выбрасывается исключение с последующей визуализацией в браузере. В процессе выполнения загрузки схемы Система обновляет элементы графического интерфейса, информируя пользователя о процессе загрузки (отображается число загруженных и оставшихся для загрузки объектов).

8) Система сохраняет загруженную схему в памяти модели рабочего пространства пользователя.

9) Система выполняет автоматическое сохранение файлов обновленного рабочего пространства пользователя.

10) В соответствии с результатами выполнения выбранного пользователем действия Система обновляет виджеты графического интерфейса, перерисовывая изменения дерева рабочего пространства и формы визуализации источника.

11) Возврат к шагу 1 или к пункту 1.3.8, шагу 6.

1.3.8.3 Создание и редактирование куба с показателями

Обобщённый алгоритм работы модуля в части создания и редактирования куба с показателями может быть представлен как последовательное выполнение следующих действий:

1) Пользователь выбирает действие «Создание и редактирование куба с показателями».

2) Система подготавливает и загружает в браузер виджеты редактирования источника данных в виде набора ресурсов (шаблоны JS, CSS, HTML5).

3) Если производится редактирование существующего куба, Система инициализирует виджеты рабочей области редактирования куба заданными параметрами куба (источники, показатели, связи источников и показателей, срезы), если нет – производится инициализация виджетов значениями по умолчанию.

4) Система ожидает от пользователя одно из следующих действий:

4.1) Добавление или удаление объекта источника. Добавление источника производится в рабочем пространстве монитора данных путем перетаскивания объекта источника данных (таблицы) в дереве рабочего пространства на свободное место рабочей области виджета редактирования куба. Удаление производится нажатием на соответствующий элемент виджета источника.

4.2) Связывание или изменений связей источников и показателей/полей куба, изменение названий показателей. Связывание элемента источника (столбца таблицы) с кубом производится манипулятором типа «мышь» – пользователь выбирает нужный элемент источника и связывает его с новым или существующим показателем.

4.2.1) Система анализирует тип данных связанного объекта источника и формирует поле куба. Если связывание производится с существующим показателем, Система формирует общее поле/показатель, использующийся для связывания нескольких источников (INNER JOIN) по данному полю. Если с новым – Система создает новый показатель/поле куба.

4.2.2) Система обрабатывает сообщение с действием пользователя по изменению названия поля, модифицируя и обновляя модель рабочего пространства пользователя (шаги с 5 по 7).

4.3) Создание или редактирование срезов

4.3.1) Система обрабатывает сообщение с действием пользователя по добавлению нового среза, модифицируя и обновляя модель рабочего пространства пользователя (шаги с 5 по 7). Если выполняется редактирование существующего среда, шаг пропускается.

4.3.2) Система обрабатывает сообщение с действием пользователя по редактированию среза – подготавливает и загружает в браузер виджеты редактирования среза в виде набора ресурсов (шаблоны JS, CSS, HTML5).

4.3.3) Система визуализирует форму редактирования среза. Если производится редактирование существующего среза, Система инициализирует виджеты актуальными параметрами среза (название, запрос к кубу, значения параметров запроса), если создается новый – параметрами по умолчанию.

4.3.4) Система обрабатывает от пользователя изменения параметров среза, после чего обрабатывает и выполняет запрос к кубу (шаг 4.4). В случае ошибки запроса формируется и выбрасывается исключение с последующей визуализацией в браузере путем формирования сообщения с описанием ошибки.

4.4) Предварительная визуализация данных источника, куба или среза.

4.4.1) Система ожидает от пользователя сообщение с действием визуализации данных (пользователь кликает по объекту на рабочей области редактирования куба).

4.4.2) Система формирует запрос к кубу, преобразует его к формату источника, выполняет соединение с источником и загружает данные в систему визуализации на сервер обработки данных ПО «Подсистема аналитики и визуализации результатов семантической интеграции разнородных ИР».

4.4.3) Система преобразует данные в формат JSON, загружает данные в браузер пользователя и выполняет визуализацию данных в табачном виджете. При использовании прокрутки таблицы данные автоматически дозагружаются. Объект

соединения с источником сохраняется до момента выполнения пользователем следующего запроса.

5) Система сохраняет настройки куба в памяти модели рабочего пространства пользователя.

6) Система выполняет автоматическое сохранение файлов обновленного рабочего пространства пользователя.

На диске рабочее пространство сохранено в виде двух типов файлов:

– файл рабочего пространства в формате JSON;

– файл ресурса рабочего пространства пользователя в формате JSON с параметрами куба.

7) В соответствии с результатами выполнения выбранного пользователем действия Система обновляет виджеты графического интерфейса, перерисовывая изменения дерева рабочего пространства и формы рабочего пространства редактирования куба.

8) Возврат к шагу 1 или к пункту 1.3.8, шагу 6.

1.3.8.4 Создание и редактирование визуализации

Обобщённый алгоритм работы модуля в части создания и редактирования визуализации с показателями может быть представлен как последовательное выполнение следующих действий:

1) Пользователь выбирает действие «Создание и редактирование куба с показателями».

2) Система подготавливает и загружает в браузер виджеты редактирование источника данных в виде набора ресурсов (шаблоны JS, CSS, HTML5).

3) Если производится редактирование существующего куба, система инициализирует виджеты рабочей области редактирования куба заданными параметрами куба (источники, показатели, связи источников и показателей, срезы), если нет – производится инициализация виджетов значениями по умолчанию.

4) Система ожидает от пользователя одно из следующих действий:

4.1) Размещение виджета диаграммы на рабочей области редактора визуализации

4.1.1) Система обрабатывает от пользователя сообщение о размещении виджета на рабочей панели визуализации. Размещение виджета производится в рабочем пространстве редактора визуализации монитора данных путем перетаскивания объекта виджета из соответствующей вкладки панели виджетов на поле рабочей области визуализации. Удаление производится нажатием на соответствующий элемент виджета.

4.1.2) Система выполняет отображение виджета на панели в режиме редактирования.

4.1.3) Система извлекает перечень выходных полей среза, выполняет их отображение в редакторе виджета, модифицируя и обновляя модель рабочего пространства пользователя (шаги с 5 по 7).

4.2) Настройка виджета диаграммы.

4.2.1) Система обрабатывает от пользователя сообщение об открытии визуализации для редактирования.

4.2.2) Задание источника данных (среза).

4.2.2.1) Система обрабатывает полученное от пользователя сообщение об установлении источника данных для виджета. Задание источника производится в рабочем пространстве редактора виджета монитора данных путем перетаскивания объекта источника данных (среза) в дереве рабочего пространства на поле источника данных редактора виджета. Удаление производится нажатием на соответствующий элемент виджета.

4.2.2.2) Система извлекает перечень выходных полей среза, выполняет их отображение в редакторе виджета, модифицируя и обновляя модель рабочего пространства пользователя (шаги с 5 по 7).

4.2.3) Настройка полей для отображения.

4.2.3.1) Система обрабатывает полученное от пользователя сообщение о переключении виджета в режим редактирования. Если действие выполняется для

нового виджета, шаг пропускается – режим редактирования включается автоматически.

4.2.3.2) Система обрабатывает полученные от пользователя сообщения о задании или редактирования полей виджета. Редактирование полей выполняется путем определения значений для соответствующих полей виджета одним из предложенных значений или заданием самого значения в поле ввода (константы).

4.2.3.3) Система модифицирует и обновляет модель рабочего пространства пользователя (шаги с 5 по 7).

4.2.3.4) Для каждого поля повторение действий с 4.2.2.2 по 4.2.2.4.

4.2.4) Настройка параметров виджета

4.2.4.1) Система обрабатывает полученные от пользователя сообщения об изменении параметров настройки виджетов. Редактирование параметров выполняется путем определения значений для соответствующих полей виджета одним из предложенных значений или заданием самого значения в поле ввода (константы).

4.2.4.2) Система модифицирует и обновляет модель рабочего пространства пользователя (шаги с 5 по 7).

4.2.4.3) Для каждого поля повторение действий с 4.2.3.1 по 4.2.3.3.

4.2.5) Тестовое отображение визуализаций.

4.2.5.1) Система обрабатывает полученное от пользователя сообщение переключения визуализации (всех или конкретного виджета) в режим отображения.

4.2.5.2) Система производит отображение визуализации аналогично пункту 1.3.8, шаги с 5 по 9.

5) Система сохраняет настройки визуализации в памяти модели рабочего пространства пользователя.

6) Система выполняет автоматическое сохранение файлов обновленного рабочего пространства пользователя. На диске рабочее пространство сохранено в виде двух типов файлов:

– файл рабочего пространства в формате JSON;

– файл ресурса рабочего пространства пользователя в формате JSON с настройками визуализации.

7) В соответствии с результатами выполнения выбранного пользователем действия Система обновляет виджеты графического интерфейса, перерисовывая изменения дерева рабочего пространства и формы рабочего пространства редактора визуализации.

8) Возврат к шагу 1 или к пункту 1.3.8, шагу 6.

1.3.9 Описание алгоритма работы подсистемы обратной связи

Функциональность ПС ОС реализована в составе подсистемы ввода данных (ПС Ввода). Эта функциональность предусматривает в случае, когда в процессе диалога с пользователем подсистемы у пользователя возникли вопросы к предоставляемой Системой информации, следующие действия пользователя:

– зафиксировать состояние интерфейса путём создания скриншота окна с диалогом,

– написать сопровождающее сообщение,

– отправить сообщение вместе с зафиксированным состоянием интерфейса группе поддержки на специальный адрес.

Группа поддержки по результатам анализа полученного сообщения выполняет корректирующие мероприятия.

1.4 Ограничения, накладываемые на область применения программы

Система может быть использована в рамках её назначения, будучи развернутой на соответствующей программно-аппаратной платформе и настроенной согласно документу АВКС.00223-01 32 01 «Система поддержки принятия врачебных решений с использованием технологий искусственного интеллекта для контроля (оценки) качества лечения в соответствии с утвержденными клиническими рекомендациями (СППВР ИИ). Руководство системного программиста».

Система предназначена для работы под управлением ОС Ubuntu 20.04 LTS.

2 УСЛОВИЯ ПРИМЕНЕНИЯ

2.1 Требования к техническим (аппаратным) средствам

Система функционирует на технических средствах, имеющих следующие характеристики:

1) рабочие станции для приложения-клиента, предназначенные для обеспечения пользователей и доступа к серверам, должны иметь:

– процессоры с частотой не менее 2 ГГц;

– объем оперативной памяти не менее 8 Гб;

– видеокарты с поддержкой OpenGL (при необходимости обеспечения возможности работы с графовым представлением данных);

2) сервер баз данных, совмещенный с сервером приложений:

– серверная платформа, совместимая с x86-64, включающая:

– процессор Intel Xeon E5/E7 от v3 до v5 с восемью ядрами и тактовой частотой не менее 2 ГГц;

– накопители на жестких магнитных дисках не менее 300 Гбайт SAS HDD 15000 rpm, не менее 2-х шт.;

– оперативная память – от 128 Гб;

– сетевой контроллер, совместимый с Fast Ethernet 1000BASE-TX;

– источник бесперебойного питания;

3) локальная вычислительная сеть, предназначенная для обеспечения взаимодействия серверов и рабочих станций, средств резервного копирования и хранения резервных копий.

Серверы должны подсоединяться к локальной вычислительной сети Ethernet посредством типового коммуникационного оборудования. При необходимости получения информации из внешних источников в сети Интернет серверы должны иметь возможность доступа к глобальной вычислительной сети Интернет через локальную вычислительную сеть Ethernet.

2.2 Требования к программным средствам (другим программам)

Система функционирует на общем программном обеспечении (ОПО), описание которого приведено в таблице 1.

Таблица 1 - Характеристики ОПО

Общее программное обеспечение	Вид общего программного обеспечения	Программный продукт
Серверное ПО	Операционная система (ОС)	Linux Ubuntu 20.04 LTS
	Виртуальная машина Java	JRE 1.8
	Веб-браузер	FireFox последней версии

На технологическом уровне для функционирования системы используется трехзвенная клиент-серверная архитектура. Компоненты трехзвенной архитектуры, с точки зрения программного обеспечения, реализуют сервер БД, сервер приложений и приложение-клиент.

Сервер БД представлен сервером БД PostgreSQL 9.4; в качестве HTTP-сервера на сервере приложений используется библиотека Jetty (свободный контейнер сервлетов, написанный на Java); роль клиента выполняет веб-браузер FireFox.

2.3 Общие характеристики входной и выходной информации

Входная информация может быть отнесена к двум категориям:

– информация, необходимая для формирования ГБД;

– информация, вводимая пользователем в процессе диалога с системой с целью реализации основной функции системы – получения из ГБД релевантной состоянию пациента информации.

Источниками информации первого типа являются массивы специальной медицинской информации: медицинские библиотеки и сформированные на их основе базы знаний и примыкающие к ним терминологические системы. Для извлечения, идентификации и связывания данных из некоторых источников могут применяться методы семантической обработки текстовой информации.

Структурированные результаты обработки источников первого типа используются для формирования ГБД. За обработку информации первого типа отвечают ПС Сбора, ПС Извлечения, ПС Ведения НСИ. ПС Обучения играет вспомогательную роль, обеспечивая функциональность ПС Извлечения и ПС Сбора.

Информация второго типа вводится пользователем в процессе диалога с ПС Ввода, которая обеспечивает идентификацию вводимых пользователем данных относительно используемого терминологического словаря, лежащего в основе ГБД. К информации второго типа относятся:

– симптомы;

– синдромы;

– диагнозы всех типов (ранее поставленные и подтверждённые, хронические, предполагаемые в данный момент);

– методы диагностики, проведённая диагностика, результаты проведённой диагностики:

– методы лечения, применяемые в данный момент или планируемые.

Запросная система ГБД или система вывода на основании введённых пользователем данных получает из ГБД:

– вероятные диагнозы;

– рекомендованные методы диагностики;

– рекомендованное лечение.

Таким образом, перечисленные выше полученные из ГБД данные, представленные пользователю, являются выходными.

2.4 Требования и условия организационного характера

Персонал, принимающий участие в эксплуатации программы, подразделяется на следующие категории специалистов:

– персонал, осуществляющий информационно-аналитическую деятельность, формулирующий постановки информационно-аналитических задач;

– персонал, реализующий постановки информационно-аналитических задач, в том числе в части поиска источников и подготовки исходных данных, их получения и загрузки в систему, настройки аналитической обработки данных, визуализации данных и результатов расчетов;

– персонал, занимающийся техническим обслуживанием системы, в том числе системным администрированием.

Информационно-аналитическая деятельность – постановка и реализация информационно-аналитических задач осуществляется предметными аналитиками с квалификацией врач-кибернетик, обладающий знаниями в лечебном деле и в информационных технологиях, их применении в медицинских системах.

Рекомендуется следующий состав эксплуатационного персонала:

– аналитик-специалист предметной области, в данном случае, врач-кибернетик, имеющий опыт работы с автоматизированными системами (программным обеспечением, базами данных) не менее 2 лет;

– аналитик-специалист по подготовке данных, их анализу, разработке витрин данных, аналитических панелей, разработке приложений баз данных и пользовательских интерфейсов;

– системный администратор, обладающий квалификацией для администрирования Linux Ubuntu и СУБД PostgreSQL.

Аналитик-специалист предметной области выполняет действия по подготовке данных, разработке и модернизации запросной системы ГБД.

Аналитик-специалист по подготовке данных и их анализу выполняет конфигурирование портала, аналитических панелей, отчётов в соответствии с инструкциями соответствующих «Руководств оператора», участвует в постановке информационно-аналитических задач.

Системный администратор поддерживает функционирование Системы, выполняет развертывание компонент Системы, необходимые настройки параметров компонент и резервное копирование/восстановление данных в соответствии с инструкциями «Руководства системного программиста», обеспечивает доступ пользователей к данным Системы.

Взаимодействие специалистов в процессе эксплуатации системы и особенно в процессе разработки и реализации новых информационно-аналитических задач должно осуществляться в единой системе управления изменениями, использующей соответствующую автоматизированную систему управления задачами. На этапе модернизации и разработки Системы в качестве системы управления задачами используется Redmine, поддерживающая общее информационное поле взаимодействующих участников проекта.

Эксплуатационный персонал должен пройти обучение у разработчика Системы в соответствии с программной обучения, подготавливаемой разработчиком Системы.

Конечный пользователь Системы является специалистом предметной области, для которой разрабатываются информационно-аналитические задачи. Для этой категории пользователей должно быть предусмотрено специальное обучение и, как минимум, специальный инструктаж.

2.5 Требования и условия технического и технологического характера

Для обеспечения требований безопасности серверное оборудование должно быть также оборудовано следующими программными и аппаратными средствами:

– программно-аппаратный замок, контролирующий загрузку ОС;

– средства антивирусной защиты;

– средства защиты от несанкционированного доступа.

Для повышения отказоустойчивости серверное оборудование Системы дублируется и объединяется в отказоустойчивые кластеры.

3 ОПИСАНИЕ ЗАДАЧИ

Компоненты Системы реализуют следующие комплексы задач:

– комплекс задач создания ГБД;

– комплекс задач предоставления информации конечному пользователю.

3.1 Комплекс задач создания ГБД

Комплекс задач создания ГБД состоит из следующих задач:

– запись исходных документов в реляционную таблицу ГБД для семантической обработки в унифицированном представлении;

– преобразование справочников из исходных представлений и запись в реляционные таблицы данных справочников в унифицированном представлении;

– обучение майнеров;

– семантическая обработка – извлечение метаданных из исходных текстовых документов и записей справочников;

– поддержка разработки схемы ГБД;

– формирование файлов для загрузки ГБД;

– загрузка ГБД.

Далее описываются содержание задач и указывается подсистема, автоматизирующая выполнение этих задач.

3.1.1 Запись исходных документов в реляционную таблицу ГБД для семантической обработки в унифицированном представлении

ПС Сбора представляет собой набор скриптов, с помощью которых Пользователь-аналитик осуществляет загрузку полученных из различных источников данных в ПС Хранения с целью их дальнейшей семантической обработки.

Могут загружаться как структурированные, так и не структурированные данные в различных форматах. Источником может быть:

– набор файлов в папке файловой системы,

– таблица или набор таблиц реляционной базы данных,

– отдельный структурированный csv-файл, содержащий текстовые и цифровые данные.

При всём разнообразии источников как по форматам представления, так и по объёмам, целевым способом и местом хранения данных ПС Сбора являются таблицы реляционной базы данных, где подлежащие семантической обработке текстовые данные размещаются в определённом формате, поддерживаемом модулем АВКС.00309-01 «Семантический индекс для PostgreSQL».

3.1.2 Преобразование справочников из исходных представлений и запись в реляционные таблицы данных справочников в унифицированном представлении

Полученные из различных источников справочники, содержащие различные типы данных, представленных в различных форматах, загружаются в ПС Хранения скриптами ПС Сбора с целью их дальнейшей семантической обработки и возможного использования для формирования ГБД.

Могут загружаться как структурированные, так и не структурированные данные в различных форматах. Источником может быть:

– набор файлов в папке файловой системы,

– таблица или набор таблиц реляционной базы данных,

– отдельный структурированный csv-файл, содержащий текстовые и цифровые данные.

Результатом обработки скриптами являются реляционные таблицы с данными в виде, удобном для формирования загрузочных файлов ГБД. Запись таких файлов и формирование ГБД – ответственность ПС Ведения НСИ.

3.1.3 Обучение майнеров

Обучение майнеров является задачей ПС обучения, которая предназначена для формирования обучающих наборов данных и тренировки нейронной сети nlp-модуля:

– создание датасета (корпуса текстов) для тренировки майнеров – нейронных сетей;

– тренировка унимодальных майнеров, использующих обученную нейронную сеть, для каждого отдельного типа объектов: «лекарственные препараты», «состояния/заболевания», «синдромы», «анатомия» и т.п.

С помощью программы пользователь может:

– создавать словники (словари-справочники);

– создавать корпуса текстов (наборы текстовых документов);

– запускать обучение майнера нейронной сети;

– выполнять оценку качества работы обученного майнера.

Обученные майнеры используются подсистемой извлечения метаданных (ПС извлечения).

3.1.4 Семантическая обработка – извлечение метаданных из исходных текстовых документов и записей справочников

Извлечение метаданных из исходных текстовых документов и записей справочников является задачами ПС Сбора и ПС Извлечения.

В рамках функциональности ПС Сбора скрипт-обработчик, входящий в состав модуля «Семантический индекс для PostgreSQL», считывает данные из таблицы реляционной базы данных, в которой в унифицированном виде хранятся исходные тексты для обработки, и записывает метаданные, полученные в результате обработки, в результирующие таблицы реляционной базы данных, формируя так называемый семантический индекс.

ПС Извлечения обладает возможностью извлекать словарные сущности из вводимых с помощью неё текстов на естественном языке, а также определять их типы.

3.1.5 Поддержка разработки схемы ГБД и запросов к ГБД

Вся подготовленная информация о разнообразных медицинских фактах (объектах и их связях) должна быть сведена в единой схеме ГБД. На основании применяемой схемы также должны быть разработаны запросы для машины вывода, посредством которых из ГБД будут извлечены данные, интересующие конечного пользователя.

Разработка данной схемы является нетривиальной задачей с учетом большого разнообразия типов объектов, связей и источников, к тому же обладающих спецификой своей собственной объектной модели. Так, внутри каждой справочной системы, например, могут существовать различные связи, выражающие похожие отношения. То же верно и относительно типов объектов разных справочных систем. Всё это сопровождается большим объёмом представляемой информации – в ГБД записываются миллионы объектов, десятки миллионов связей с учетом покрытия всего медицинского домена. Поэтому разработка схемы ГБД представляет собой отдельный затратный процесс, который для повышения его эффективности необходимо поддерживать применением отдельного инструментария.

Процесс разработки схемы ГБД и запросов поддерживается в функциональных рамках подсистем извлечения метаданных (ПС Извлечения), настройки и формирования отчетных форм (ПС ФОФ) и предоставления аналитических данных (ПС Предоставления АД).

ПС Извлечения предоставляет пользователю-аналитику разметку текстовых материалов, которые могут использоваться для извлечения отдельных объектов (и фактов), позволяет искать необходимую информацию в наборах документов. В результате этой обработки пользователь-аналитик может корректировать схему ГБД и добавлять извлечённую информацию в ГБД.

ПС ФОФ позволяет формировать аналитические панели, с помощью которых пользователь-аналитик работать с исходным содержанием справочных систем со всеми взаимосвязями и особенностями. ПС Предоставления АД отвечает за отображение аналитических панелей.

3.1.6 Формирование файлов для загрузки ГБД

Формирование специальных файлов для загрузки ГБД относится к ответственности ПС Ведения НСИ, которая представляет собой набор скриптов.

Для формирования могут быть использованы данные в таблицах реляционной БД, а также дистрибутивные файлы справочных и иных систем. При этом данные из дистрибутивных файлов справочных и других систем могут сначала загружаться в реляционную базу данных и потребляться ПС Ведения НСИ из таблиц реляционной системы. Такие таблицы в реляционной БД формируются подсистемой сбора исходных документов (ПС Сбора) в результате обработки исходных дистрибутивов медицинских библиотек, загруженных и обработанных текстовых материалов, из которых извлечена метаинформация.

Указанные скрипты и программы формируют загрузочные файлы в специальном формате загрузчика ГБД с учетом применяемой схемы ГБД, а именно: применяемых объектов, связей, свойств объектов и связей, меток объектов.

3.1.7 Загрузка ГБД

Пользователь-аналитик загружает ГБД с помощью специальной программой-загрузчика. На вход программы подаются загрузочные файлы, хранящие данные в специальном формате загрузчика с учетом применяемой схемы ГБД, а именно: применяемых объектов, связей, свойств объектов и связей, меток объектов.

После загрузки ГБД может быть непосредственно использована ПС Ввода клинических данных для предоставления экспертной информации конечному пользователю.

3.2 Комплекс задач предоставления информации конечному пользователю

Комплекс задач предоставления информации конечному пользователю решается подсистема ввода данных (ПС Ввода), которая представляет собой веб-приложение, формирующее сложные последовательности запросов к ГБД.

Перечень сценариев, инициируемых пользователем через веб-клиент:

1) Выполнить ввод симптомов или показаний.

2) Получить список симптомов, часто встречающихся с отобранными ранее;

3) Получить группированный список предполагаемых диагнозов.

4) Получить дополнительную информацию по диагнозу:

– похожие диагнозы;

– диагностические признаки;

– медикаменты для лечения;

– встречающиеся симптом.

Конечный пользователь также может быть заинтересован в поиске информации по источникам с применением кластеризации и разметки объектами по словарям. Эта функциональность поддерживается в функциональных рамках подсистем извлечения метаданных (ПС Извлечения), настройки и формирования отчетных форм (ПС ФОФ) и предоставления аналитических данных (ПС Предоставления АД).

ПС Извлечения предоставляет пользователю разметку текстовых материалов, которые могут использоваться для извлечения отдельных объектов (и фактов), позволяет искать необходимую информацию в кластеризованных наборах документов.

4 ВХОДНЫЕ И ВЫХОДНЫЕ ДАННЫЕ

4.1 Входные данные

4.1.1 Входные данные подсистемы сбора исходных документов

Входными данными ПС Сбора являются:

– тексты на естественном языке;

– структурированные или частично структурированные справочники, содержащие текстовую информацию, подлежащую обработке семантическими методами, реализованными в модуле АВКС.00309-01 «Семантический индекс для PostgreSQL».

4.1.2 Входные данные подсистемы хранения данных

ПС Хранения хранит данные всех других подсистем, позволяя им обмениваться данными между собой.

Соответственно этому, входными данными для ПС Хранения являются все данные, создаваемые всеми другими подсистемами.

4.1.3 Входные данные подсистемы ввода данных

Входными данными для ПС Ввода являются:

– данные, загружаемые из ГБД;

– данные, вводимые пользователем в элементы ввода на графическом интерфейсе ПС Ввода.

4.1.4 Входные данные подсистемы ведения нормативно-справочной информации

Входными данными ПС Ведения НСИ являются исходные данные для формирования ГБД – справочники с биомедицинской информацией, описывающие различные объекты, имеющие значение для диагностики и лечения, в том числе

содержащие информацию о связях объектов. Эти справочники, в зависимости от способа получения и использованных для их формирования методов обработки, могут храниться в таблицах PostgreSQL либо в отдельных csv-файлах в файловой системе.

4.1.5 Входные данные подсистемы извлечения метаданных

Входными данными для ПС Извлечения являются:

– данные, загружаемые из ГБД;

– данные, вводимые пользователем в элементы ввода на графическом интерфейсе ПС Извлечения.

4.1.6 Входные данные подсистемы настройки и формирования отчетных форм

Входными данными для настройки модуля «Монитор данных» – основного компонента модуля ускорения вычислений «DataMonitor», входящего в ПС ФОФ, являются:

а) данные пользователей;

б) конфигурационные данные, хранимые в специальной схеме реляционного сегмента ПС Хранения – рабочей области модуля «DataMonitor», содержащие описание:

– проекта;

– источников данных;

– кубов;

– срезов;

– визуализаций (виджетов);

– аналитических панелей;

– другая конфигурационная информация, релевантная пользователю;

в) данные, загружаемые из источников данных – таблиц реляционного сегмента ГБД, с целью их визуализации на аналитических панелях, отображаемых ПС Предоставления АД.

4.1.7 Входные данные подсистемы обучения

Входными данными для ПС Обучения являются:

– наборы данных по тематике обучения;

– словари по тематике обучения.

4.1.8 Входные данные подсистемы предоставления аналитических данных

Входными данными для ПС Предоставления АД являются:

а) данные пользователей, предоставляемые системой сквозной аутентификации Astra Linux Directory;

б) конфигурационные данные, хранимые в рабочей области модуля, содержащие описание:

– проекта;

– источников данных;

– кубов;

– срезов;

– визуализаций (виджетов);

– аналитических панелей;

– другая конфигурационная информация, релевантная пользователю;

в) данные, загружаемые из источников данных – таблиц БД, с целью их визуализации на аналитических панелях.

4.1.9 Входные данные подсистемы обратной связи

Входными данными для ПС ОС являются:

– параметры запросов, выполненных в результате диалога пользователя с ПС Ввода;

– текстовое сообщение пользователя, описывающего возникшую проблему.

4.2 Выходные данные

4.2.1 Выходные данные подсистемы сбора исходных документов

Выходными данными для ПС Сбора являются:

– таблица в реляционной БД, содержащая тексты в унифицированном представлении для семантической обработки методами модуля АВКС.00309-01 «Семантический индекс для PostgreSQL».

4.2.2 Выходные данные подсистемы хранения данных

ПС Хранения хранит данные всех других подсистем, позволяя им обмениваться данными между собой.

Соответственно этому, выходными данными для ПС Хранения являются все данные, кроме данных внешних источников, потребляемые всеми другими подсистемами.

4.2.3 Выходные данные подсистемы ввода данных

Для пользователя ПС Ввода выходными данными являются списки текстовой информации, содержащие описание клинической картины рассматриваемого заболевания, формируемые в зависимости от действий пользователя и отображаемые в интерфейсе ПС Ввода для пользователя.

4.2.4 Выходные данные подсистемы ведения нормативно-справочной информации

Выходными данными ПС Ведения НСИ являются сформированные загрузочные файлы с данными в унифицированном формате для формирования ГБД программой-загрузчиком ГБД.

4.2.5 Выходные данные подсистемы извлечения метаданных

Для пользователя ПС Извлечения формирует и отображает метаинформацию по выполненной разметке текстов:

– списки кластеров размеченных текстов;

– списки терминов, найденных в текстах, с типами терминов.

4.2.6 Выходные данные подсистемы настройки и формирования отчетных форм

Для конечного пользователя – потребителя информации выходными данными модуля «Монитор данных», основного компонента модуля ускорения вычислений «DataMonitor», входящего в ПС ФОФ, являются:

а) настроенные аналитические панели, визуализируемые монитором данных, с настроенными виджетами следующих типов:

– табличный;

– географическая карта;

– индикатор процесса;

– текстовой;

– элемент HTML;

– линейная диаграмма;

– столбчатая диаграмма;

– временной диапазон;

– круговая диаграмма;

– полярная диаграмма;

– карта дерева;

– лучевая диаграмма;

– граф;

– пенное дерево;

б) отдельные виджеты перечисленных типов, размещенные на страницах пользовательского портала;

в) выгруженные данные из табличных виджетов в форматах:

– json;

– xls;

– csv;

– изображений виджетов.

Технически, помимо перечисленных выше визуализаций (виджетов и аналитических панелей), а также выгрузок, выходными данными являются:

а) конфигурационные данные, записанные в рабочей области Монитора данных, находящейся в специальной схеме реляционного сегмента ГБД, и описывающие:

– проекты;

– источники данных для кубов;

– OLAP-кубы;

– вычислительные и фильтрационные срезы;

– виджеты;

– аналитические панели;

б) данные о действиях пользователей, записываемые в рабочую область монитора данных.

4.2.7 Выходные данные подсистемы обучения

Выходными данными для ПС Обучения являются:

– размеченные наборы данных для обучения;

– обученная модель;

– оценка качества обученной модели.

4.2.8 Выходные данные подсистемы предоставления аналитических данных

Для конечного пользователя – потребителя информации выходными данными ПС Предоставления АД являются:

а) настроенные аналитические панели, визуализируемые монитором данных, с настроенными виджетами следующих типов:

– табличный;

– географическая карта;

– индикатор процесса;

– текстовой;

– элемент HTML;

– линейная диаграмма;

– столбчатая диаграмма;

– временной диапазон;

– круговая диаграмма;

– полярная диаграмма;

– карта дерева;

– лучевая диаграмма;

– граф;

– пенное дерево;

б) отдельные виджеты перечисленных типов, размещенные на страницах пользовательского портала;

в) выгруженные данные из табличных виджетов в форматах:

– json;

– xls;

– csv;

– изображений виджетов.

Технически, помимо перечисленных выше визуализаций (виджетов и аналитических панелей), а также выгрузок, выходными данными являются:

а) конфигурационные данные, записанные в рабочей области Монитора данных, находящейся в специальной схеме реляционного сегмента ГБД, и описывающие:

– проекты;

– источники данных для кубов;

– OLAP-кубы;

– вычислительные и фильтрационные срезы;

– виджеты;

– аналитические панели;

б) данные о действиях пользователей, записываемые в рабочую область монитора данных.

4.2.9 Выходные данные подсистемы обратной связи

Выходными данными ПС ОС является сообщение группе поддержки, содержащее:

– параметры запросов, выполненных в результате диалога пользователя с ПС Ввода;

– текстовое сообщение пользователя, описывающего возникшую проблему.

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

API	–	(англ. Application Programming Interface) программный интерфейс приложения, интерфейс прикладного программирования;
HTTPS	–	(англ. HyperText Transfer Protocol Secure) расширение протокола HTTP для поддержки шифрования в целях повышения безопасности;
jdbc	–	(англ. Java DataBase Connectivity) соединение с базами данных на Java — платформенно-независимый промышленный стандарт взаимодействия Java-приложений с различными СУБД, реализованный в виде пакета java.sql, входящего в состав Java SE;
OLAP	–	(англ. Online analytical processing) Онлайн-аналитическая обработка, или OLAP, – это технология вычисления быстрых ответов на многомерные аналитические запросы. OLAP является частью более широкой категории бизнес-аналитики, которая также включает реляционные базы данных, формирование отчетов и интеллектуальный анализ данных.
PostgreSQL	–	свободная объектно-реляционная система управления базами данных;
SQL-запрос	–	(англ. Structured Query Language) «язык структурированных запросов» – язык программирования, применяемый для создания, модификации и управления данными в реляционной базе данных, управляемой соответствующей системой управления базами данных;
БД	–	база данных;
ГБД	–	гетерогенная база данных;
НД	–	набор данных;
НСИ	–	нормативно-справочная информация;
ОПО	–	общее программное обеспечение;
ОС	–	операционная система;
РБД	–	реляционная база данных;
СУБД	–	система управления базами данных.

Лист регистрации изменений

Номера листов (страниц)					Всего листов (страниц) в докум.	№ документа	Входящий № сопроводительного документа и дата	Подп.	Дата
Изм.	измененных	замененных	новых	аннулированных					