14. Параллельный генетический алгоритм отбора значимых факторов, влияющих на эволюцию сложной системы Мокшин В.В. Вестник Казанского государственного технического университета им. А.Н. Туполева. 2009. № 3. С. 89-93.

15. Выбор метрики в машинном обучении (Random forest) [Электронный ресурс]: Блог кампании Деталитика. – Режим доступа: http://blog.datalytica.ru/2018/05/blog-post.html (Дата обращения: 07.11.2021)

А.Ю. Жигалов, И.П. Болодурина, Д.И. Парфенов, Л.С. Гришина

РАЗРАБОТКА ГРАФОВОЙ МОДЕЛИ СТРУКТУРНЫХ И СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ МЕЖДУ СУЩНОСТЯМИ ДОКУМЕНТОВ ДЛЯ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

(Оренбургский государственный университет)

Сегодня наблюдается взрывной рост количества информации, создаваемой людьми и машинами на естественном языке. Аналитическое агентство IDC прогнозирует рост совокупного объема данных, накопленных человечеством, до 163 зеттабайт к 2025 году. Основной частью таких данных являются неструктурированные данные, такие как фотографии, видеозаписи, аудиозаписи, а также тексты на естественном языке. Постоянное увеличение интенсивности потока входящей текстовой информации делает все более важной задачу обработки естественного языка, в частности — русского языка.

Важнейшей проблемой является лексическая многозначность, требующая от машины понимания контекста и предметной области, в которой употребляется каждое многозначное слово [1]. Такие сведения представляются в семантических сетях — специальных высококачественных базах знаний, представляющих машиночитаемые сведения об окружающем мире в виде понятий и связей между ними. Связи между понятиями задают семантическую иерархию, которая позволяет решать различные задачи машинного понимания естественного языка.

В настоящее время обработка естественного языка (Natural language processing, NLP) является наиболее инновационным направлением искусственного интеллекта. При решении многих задач NLP, таких, как распознавание и синтез речи, машинный перевод, классификация текстов, разработка диалоговых систем, в последнее время достигнут значительный прогресс на основе нейросетевых методов машинного обучения [2]. В первую очередь, исследователи занимаются решением универсальных задач, которые могут найти применение в различных областях таких как финансы, медицина, медиа и реклама. К таким задачам можно отнести генерацию продолжения текста (сети GPT-2,

GPT-3, T5 YaLM), поиск ответа на вопрос по тексту, выделение именованных сущностей (на основе разных версий BERT).

Применение в современных работах классических статистических моделей и методов для задач семантического анализа текстов и информационного поиска (например, TF-IDF, LSA) не позволяет решить проблему существенного различия лексического состава текстов, ограниченной лексики, используемой в документах, и лексического состава текстов. Данную проблему можно решить, реализовав семантическое сопоставление текстов, используя принципы дистрибутивной семантики на основе современных нейросетевых моделей языка word2vec, fastText, обучаемых без учителя на больших текстовых корпусах. Данные методы показывают свою эффективность в задачах определения семантической близости и разрешения лексической многозначности, в том числе и для русского языка, что подтверждается результатами соревнований в рамках семинара RUSSE [3, 4].

Целью данной работы является разработка графовой модели структурных и семантических отношений между сущностями различных слабоструктурированных документов информационных систем, необходимых для интеллектуальной обработки больших данных, на примере электронных медицинских карт пациентов.

Автоматизированные медицинские информационные системы позволяют быстро и эффективно наладить электронный документооборот, выстраивать работу с пациентом, вести оперативный учет работы административного персонала, контролировать все организационные и финансовые вопросы. Пример Оренбургской МИС представлен на рисунках 1-2.

Рис. 1. Шаблон дневника пациента на приеме

В протоколе дневника пациента 3 части: данные, анамнез жизни и данные объективного исследования. На первой странице заполняются жалобы больного, анамнез заболевания, диагноз и план обследования.

Рис. 2. Пример заполнения данных из дневника пациента

Анамнез жизни содержит уже информацию о наследственности, вредных привычках и тд. В рамках данной работы предполагается разработка модуля автоматического преобразования xml-документа и вытягивания основной информации посредством рекурсивного обхода и анализа всех веток разметки в единую графовую модель, характеризующую взаимоотношения между медицинской организацией, пациентом, случаем лечения и посещениями МО.

Рис. 3. Модуль DictParseModule автоматического преобразования xml-документа в единую графовую модель

В основе модуля DictParseModule выделения информации разношаблонных xml-протоколов лежит подход к рекурсивному перебору узлов xml, с последовательным анализом наличия содержимого (рис. 3). Отличительной особенностью предлагаемого подхода - является создание дерева записи оказанной услуги в МО, позволяющей проанализировать взаимосвязь некоторых факторов внутри документа.

Таким образом, в рамках данной работы построена графовая модели структурных и семантических отношений между сущностями различных слабоструктурированных документов информационных систем, необходимых для интеллектуальной обработки больших данных, на примере электронных медицинских карт пациентов (рис. 4).

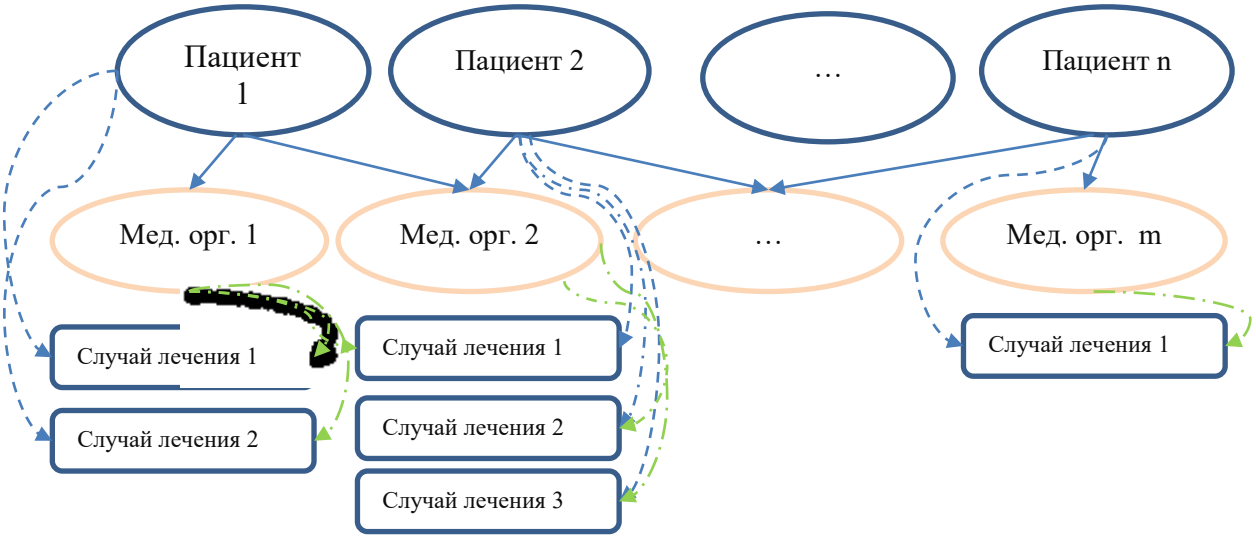

Рис. 4. Графовая модель структурных и семантических отношений МИС

Дальнейшее направление исследования включает разработку интеллектуального алгоритма автоматического выделения основных терминов в русскоязычном тексте на основе нейросетевых методов суммаризации, основанных на трансферном обучении предобученных моделей.

Исследование выполнено при финансовой поддержке РФФИ (проект № 20-07-01065) и гранта Президента Российской Федерации для государственной поддержки молодых российских ученых - кандидатов наук (№ МК-258.2022.1.6), а также стипендии Президента Российской Федерации молодым ученым и аспирантам (№ СП-919.2022.5).

Литература

1. Пикалёв Я. С. Обзор архитектур систем интеллектуальной обработки естественно-языковых текстов // Проблемы искусственного интеллекта. – 2020. – №. 4(17) – С. 45–68.

2. Батура Т. Методы автоматической классификации текстов. Международный журнал Программные продукты и системы. – 2017. – Т. 23. – С. 85–99.

3. Arefyev, N.V. Evaluating Three Corpus based Semantic Similarity Systems for Russian / N.V. Arefyev, A.I. Panchenko, A.V. Lukanin // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2015 (Moscow, RGGU) . – 2015. – Vol. 2. – P. 106–118.

4. A. Panchenko A. The First Workshop on Russian Semantic Similarity // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2015 (Moscow, RGGU). – 2015. – Vol. 2. – P. 89–105.

А.И. Иванов

СИНТЕЗ НОВЫХ ИНТЕГРО-ДИФФЕРЕНЦИАЛЬНЫХ СТАТИСТИЧЕСКИХ КРИТЕРИЕВ И ЭКВИВАЛЕНТНЫХ ИМ ИСКУССТВЕННЫХ НЕЙРОНОВ ДЛЯ МАЛЫХ ВЫБОРОК

(АО «Пензенский научно-исследовательский электротехнический институт»)

К сожалению, созданные в 20 веке статистические критерии для проверки гипотез нормальности или равномерности данных, ориентированы на большие выборки. В биометрии, медицине, экономике, во многих случаях, приходится работать с малыми выборками, например, в 16 опытов. В связи с этим возникла задача параллельного использования множества разных статистических критериев [1, 2] при анализе одной малой выборки. Каждому классическому критерию создается эквивалентный нейрон. Чем больше статистических критериев (эквивалентных искусственных нейронов) используется, тем выше достоверность, принимаемых решений, обобщающей сетью искусственных нейронов. При этом желательно синтезировать новые статистические критерии. Для синтеза может быть использован классический критерий Джинни, тогда новые критерии создаются дифференцированием входных случайных данных.

Таблица 1. Классический критерий Джинни и три его новых дифференциальных аналога

1	$D = \int_{-\infty}^{\infty} \left\| \widetilde{P}(x) - P(x) \right\| dx$	P_{EE}= 0.423	corr(D,dD) = -0.029
2	$dD = \int_{-\infty}^{\infty} \left\| \widetilde{p}(x) - p(x) \right\| dx$	P_{EE}= 0.039	corr(D,d^2D) =-0.0047
3	$d^2D = \int_{-\infty}^{\infty} \left\| \frac{d(\widetilde{p}(x))}{dx} - \frac{d(p(x))}{dx} \right\| dx$	P_{EE}= 0.047	corr(D,d^3D) =0.0367
4	$d^3D = \int_{-\infty}^{\infty} \left\| \frac{d^2(\widetilde{p}(x))}{(dx)^2} - \frac{d^2(p(x))}{(dx)^2} \right\| dx$	P_{EE}= 0.082	corr(dD,d^2D)=0.894 corr(dD,d^3D)=0.805 corr(d^2D,d^3D)=0.888
	.....		