ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
2024. Т. 6, № 4 (19). С. 111–121 http://siit.ugatu.su
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
Научная статья

УДК 004.912, 004.891.2 DOI 10.54708/2658-5014-SIIT-2024-no4-p111 EDN OTVTXR

Медицинская рекомендательная система на основе автоматического извлечения знаний из текстов

Е. А. Коровин • С. А. Чиглинцева • Е. Ю. Сазонова • О. Н. Сметанина

Аннотация. Ограничение времени приема врачей, требующее принятия решений в реальном времени, сопровождается регулярными изменениями клинических рекомендаций, основанных на новых исследованиях. Поэтому актуальность проблемной ситуации обусловлена необходимостью информационной поддержки медицинского работника с целью снижения времени поиска рекомендаций для постановки диагноза и назначения лечения пациентов с различными заболеваниями. Современное развитие инструментальных средств анализа текста позволяет автоматизировать процесс поиска рекомендаций. В статье предлагается подход к построению рекомендательной системы, способной эффективно осуществлять поиск информации о корректной постановке диагноза, лечении и профилактике заболеваний. В процессе исследования проведен анализ современного состояния проблемной ситуации, в частности, подходы к созданию рекомендательных систем в медицине. Для поиска и отбора медицинских рекомендаций выбрана, как наиболее рациональная, модель дистрибутивной семантики Word2vec, разработан алгоритм в виде комбинации поиска совпадающих предложений по регулярным выражениям и модели поиска дистрибутивно-семантических связей Word2vec.Выявлены ограничения в выборе решений для предобработки и генерации медицинских клинических рекомендаций, разработан алгоритм, способный генерировать индивидуальные рекомендации для пациентов на основе анализа совпадающих предложений по регулярным выражениям. Авторами статьи предложены архитектура и структура базы данных для хранения информации о пациентах, записях приема и диагнозах. Продемонстрировано, что созданная рекомендательная система с интегрированным алгоритмом дистрибутивной семантики Word2vec сокращает время, затрачиваемое медицинским персоналом на поиск справочной информации, и может быть полезной для специалистов медицинской отрасли. Предложенные решения имеют практическую ценность и могут служить основой для дальнейших исследований и развития автоматизированных систем в медицине.

Ключевые слова: рекомендательные системы в медицине; машинное обучение; регулярные грамматики; обработка естественного языка; методы анализа текстовых данных; дистрибутивная семантика; бронхолегочные заболевания.

Введение

Актуальность проблемной ситуации обусловлена необходимостью своевременной и правильной диагностики и лечения заболеваний. Это связано, прежде всего, с ограничением времени, отведенного на прием врачей, регулярными изменениями текстов клинических рекомендаций, вызванных новыми исследованиями. Возможность автоматизировать процесс анализа текста и строить на результатах анализа рекомендации с использованием современных моделей и методов также подтверждает актуальность решаемой задачи.

Вопросами применения цифровых технологий и, в частности, разработкой и анализом рекомендательных медицинских систем занимались В. В. Грибова, Е. А. Шалфеева [1], И. П. Болодурина [2], В. В. Цурко [3], Б. А. Кобринский [4], А. Гусев, Т. Зарубина [5].

Авторы работ [6, 7] приводят обзоры медицинских рекомендательных систем, в [8] представляют классификацию рекомендательных систем (фильтрация содержимого, коллаборативная фильтрация и гибридные системы). Дж. Г. Д. Очоа, О. Чишар и Т. Шимпер [9] добавляют такой аппарат, как непрерывнозначная логика и операторы принятия многокритериальных решений.

Несмотря на то, что данными вопросами занимаются многие специалисты, применение новых моделей и методов или комбинации методов может повысить эффективность решения.

В статье отражены результаты анализа современного состояния проблемной ситуации, предложена концепция для поиска и отбора медицинских рекомендаций на основе модели дистрибутивной семантики Word2vec и регулярных выражений; выявлены ограничения в выборе решений для предобработки и генерации медицинских клинических рекомендаций, разработаны алгоритмы для поиска индивидуальных рекомендаций для пациентов на основе анализа совпадающих предложений по регулярным выражениям. Авторами статьи предложены архитектура СППР и структура базы данных для хранения информации о пациентах, записях приема и диагнозах.

Сравнительный анализ подходов к поиску рекомендаций

Поиск решений, соответствующих симптомам пациента в тексте клинических рекомендаций, сводится к задаче определения смысловой и контекстной близости слов.

Для решения используется набор предложений в качестве входных данных ограниченного объема, что исключает возможность использования глубоких нейронных сетей. Для оценки возможности использования современных подходов, применяемых для такого рода задач, рассмотрены поиск по регулярным выражениям, генеративный подход и интеллектуальный поиск.

Метод поиска с применением регулярных выражений основан на формальных правилах. Регулярные выражения в виде последовательности символов-шаблонов позволяют осуществлять поиск и извлечение информации путем сопоставления шаблона с текстом. Метод полезен для выполнения простых операций (извлечение шаблонных данных или поиск слов / фраз), соответствующих определенным критериям. Такой аппарат для обработки текстовых данных использовали многие специалисты [10, 11] и др.

Генеративный подход основан на моделировании вероятностных распределений для генерации новых данных, схожих с обучающим набором, в том числе и для текстовых данных – рекомендаций. Рекомендательные системы с генеративным поиском рассмотрены в работе [12]. В задачах генерации текста модели, как правило, обучаются на большом объеме текстовых данных. Для моделирования зависимостей между словами и генерации последовательностей текста часто используются рекуррентные нейронные сети, такие как LSTM (Long Short-Term Memory).

Интеллектуальный поиск может быть представлен комбинацией методов машинного обучения и искусственного интеллекта по обработке поисковых запросов и предоставления релевантных результатов. Подход часто используется в системах поиска информации для нахождения наиболее подходящих документов или ресурсов на основе запроса пользователя, включая широкий спектр моделей и методов (модели ранжирования (BM25, TF-IDF), нейронные сети (сверточные нейронные сети (CNN) и трансформеры)).

Для задачи дистрибутивной семантики, фокусирующейся на смысловой близости слов [13], может быть использована модель из семейства моделей Word2vec, основанная на архитектуре Skip-gram.

Методы извлечения таксономических отношений из текстов, основанные на шаблонах [14], обладали низкой полнотой, так как требовали определенных конструкций в ограниченном числе предложений. Однако появление векторных представлений слов [15], эмбеддингов, предоставило новые возможности для извлечения знаний из текстов. Векторные представления формируются из контекстов, в которых слова упоминаются, и сходство контекстов приводит к сходству векторных представлений слов. Это позволяет автоматически определять семантическую близость слов на основе текстовых коллекций и значительно повышает точность извлечения таксономических отношений [16].

Каждый из подходов обладает своими уникальными преимуществами и ограничениями в контексте задач извлечения информации из текста и обработки естественного языка. Так,

генеративный подход, в отличие от интеллектуального поиска и применения регулярных выражений, не обладает интерпретируемостью, что принципиально в области медицины. Также для него характерны значительные минимальные размеры корпуса текстов (более 1000000 рекомендаций), в отличие от двух других подходов (размеченный корпус от 1000 рекомендаций). По параметру «Время поиска решения» более эффективными также являются подходы с использованием интеллектуального поиска и с использованием регулярных выражений (менее секунды) и две-три секунды для генеративного подхода. Точность для генеративного подхода зависит от размера датасета и нередко не достигает минимальной.

Предлагаемое решение: модели, методы и алгоритмы для поиска рекомендаций

Предлагаемый подход заключается в применении методов, способных дать хорошие результаты при ограниченных ресурсах.

Метод Word2vec позволяет представлять слова в виде числовых векторов на основе их контекста. Этот метод использует две матрицы: матрицу W для центральных слов и матрицу D для контекстных слов $W, D \in \mathbb{R}^{Vocab \times EmbSize}$. Размерности этих матриц зависят от размера словаря и размерности векторов. Обучение модели Word2vec основано на методе максимального правдоподобия (2).

$$\sum_{CenterW_i} \mathrm{P}(\mathrm{Ctx}W_{-2}, \mathrm{Ctx}W_{-1}, \mathrm{Ctx}W_{+1}, \mathrm{Ctx}W_{+2} | \mathrm{CenterCtx}W_i; \mathrm{W}, \mathrm{D}) \to \max_{W,D}. \quad (2)$$

Метод Word2Vec стремится предсказать распределение соседних слов в заданном окне, учитывая параметры модели. Для этого слова в окне представляются через произведение более простых распределений. Эти распределения определяют вероятность встретить контекстное слово рядом с центральным (3). Для моделирования категориального распределения, которое является дискретным и принимает значения из фиксированного набора, используется функция softmax (4).

$$\mathrm{P}(\mathrm{Ctx}W_{-2}, \mathrm{Ctx}W_{-1}, \mathrm{Ctx}W_{+1}, \mathrm{Ctx}W_{+2} | \mathrm{CenterCtx}W_i; \mathrm{W}, \mathrm{D}) = \prod_j P(CtxW_i | CenterW_i; W, D), \quad (3)$$

$$P(CtxW_i | CenterW_i; W, D) = \frac{e^{w_i . d_j}}{\sum_{j=1}^{|V|} e^{w_i . d_j}} = softmax \simeq \frac{e^{w_i . d_j^+}}{\sum_{j=1}^{k} e^{w_i . d_j^-}}, k \ll |V|. \quad (4)$$

Сходство слов в модели Word2Vec вычисляется через скалярное произведение векторов центрального и контекстного слов. Однако вычисление полного softmax по всем словам словаря может быть вычислительно неэффективным. Вместо этого используется метод отрицательного сэмплирования (negative sampling), который позволяет сократить количество вычислений softmax путем выбора случайных слов. Таким образом, обучение Word2Vec сводится к обучению классификатора, который предсказывает, могут ли два слова встретиться вместе.

Идея Skip-gram в Word2Vec заключается в предсказании контекстных слов на основе центрального слова. Этот подход эффективно моделирует взаимосвязи между словами и позволяет использовать полученные векторы для различных задач обработки естественного языка, таких как кластеризация слов, поиск похожих слов и машинный перевод. Обучение Word2Vec модели оптимизируется путем минимизации функции потерь с помощью стохастического градиентного спуска.

Исследования и анализы показывают, что модель Word2Vec успешно захватывает семантические и синтаксические аспекты слов [17]. Она способна обучаться на больших объемах текстовых данных и порождать семантически богатые векторные представления слов. В связи с этим Word2Vec становится одним из самых популярных методов для работы с естественным языком и извлечения смысловой информации из текстов.

Предобработка данных в корректировке ошибок и проблем некорректного форматирования документов из формата PDF в ручном режиме на этапе исследования значительно повышает точность и надежность результатов исследований, несмотря на затраты времени.

Разработанный скрипт Preprocessing.ipynb включает три основных элемента: предварительная обработка данных, обучение модели Word2vec и поиск рекомендаций.

В процессе предварительной обработки требуется нормализовать текст рекомендаций и сохранить их во фрейм данных. В качестве входных данных поступает file.txt-документ, содержащий клинические рекомендации.

Алгоритм предобработки текста:

Шаг 1. Загрузка модулей и текста из файла file.txt.

Шаг 2. Разделение текста на предложения.

Шаг 3. Нормализация слов в каждом предложении.

Шаг 4. Перевод представления предложения в список нормализованных слов.

Шаг 4. Создание двух фреймов данных df: с двумя столбцами: «Предложения» и «Нормализованный текст» (рис. 1, *а*), со столбцом «Предложения» (рис. 1, *б*).

Шаг 5. Сохранение фреймов данных под именами «fileFull.csv» и «filePart.csv».

	Предложения	Нормализованный текст
0	Хроническая обструктивная болезнь легких (ХОБЛ...	[хронический, обструктивный, болезнь, лёгкий, ...
1	Обострения и коморбидные состояния являются не...	[обострение, коморбидный, состояние, являться,...
2	Курение остается основной причиной ХОБЛ.	[курение, оставаться, основной, причина, хобл]
3	По некоторым оценкам, в индустриальных странах...	[по, некоторый, оценка, индустриальный, страна...
4	В развивающихся странах существенное повреждаю...	[в, развивающийся, страна, существенный, повре...
5	Этиологическую роль также могут играть професс...	[этиологический, роль, также, мочь, играть, пр...

а

	Предложения
0	Хроническая обструктивная болезнь легких (ХОБЛ...
1	Обострения и коморбидные состояния являются не...
2	Курение остается основной причиной ХОБЛ.
3	По некоторым оценкам, в индустриальных странах...
4	В развивающихся странах существенное повреждаю...

б

Рис. 1 Фреймы данных df:
а – фрейм данных fileFull.csv; *б* – фрейм данных filePart.csv.

Для поиска рекомендаций использована методика «обогащения» знаний, в основе которой лежит поиск прямых совпадений по исходному пользовательскому запросу. Затем список рекомендаций расширяется за счет дополнительных вариантов, определенных на основе смысловой близости слов. Подход позволяет эффективно увеличить число рекомендаций, особенно в случаях, когда исходный запрос возвращает небольшое количество результатов.

Для генерации дополнительных рекомендаций используется обученная модель Word2Vec архитектуры Skip-gram. Модель обучается на больших текстовых наборах данных, выявляя контекстуальные связи между словами. В результате, модель строит векторные представления, где семантически близкие слова имеют близкие векторы. Таким образом, при запросе

пользователя модель может искать семантически близкие слова для расширения списка рекомендаций. Эти методы позволяют эффективно расширять перечень рекомендаций на основе смысловой близости слов и контекстуальных связей в текстовых данных.

На этапе поиска входными данными будут являться диагноз, степень его тяжести и дополнительные сведения, а в качестве выходных данных будут рекомендации.

Алгоритм поиска рекомендаций:

Шаг 1. Приведение запроса пользователя к виду списка нормализованных слов.

Шаг 2. Поиск прямых совпадений по имеющемуся корпусу рекомендаций.

Если обнаружены совпадения, то совпадения сохраняются в виде множества потенциальных рекомендаций.

Шаг 3. Поиск «близких» по смыслу и контексту слов по диагнозу, с использованием обученной модели.

Шаг 4. Отбор рекомендаций, содержащих «близкие» к исходному диагнозу слова.

Шаг 5. Объединение множества потенциальных рекомендаций и множества рекомендаций по контексту.

Программная реализация предлагаемого решения

Созданное авторами программное решение обладает широким спектром функциональных возможностей (рис. 2), включая авторизацию пользователей, возможность осуществления поиска пациентов и добавления новых записей в базу данных, а также изменение имеющейся информации о пациентах.

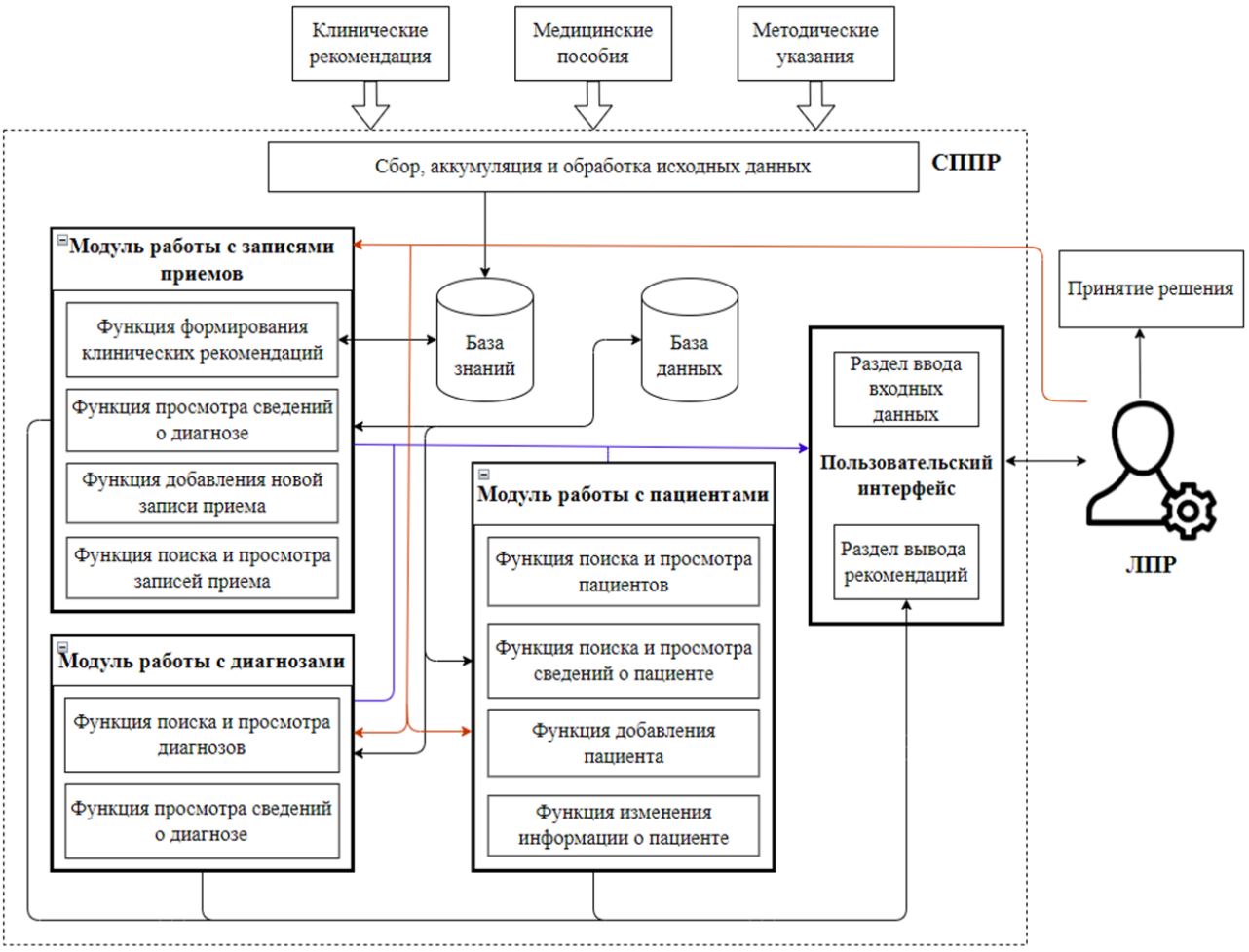

Рис. 2 Структурная схема системы поддержки принятия решений при работе приложения.

Также система позволяет осуществлять поиск и просмотр диагнозов и записей приема, добавление новых записей приема и генерацию текстовых рекомендаций. Редактирование формы с клиническими рекомендациями и сохранение рекомендаций после внесения изменений в буфер обмена также доступны в данной системе. Программа для анализа и генерации текста состоит из следующих функций: normalize_word (word) нормализует слово при помощи модуля pymorphy2 и извлекает нормальную форму слова для дальнейшей обработки; process_sentence (sentence) обрабатывает предложение, которое токенизируется при помощи nltk.word_tokenize, затем каждое слово нормализуется при помощи normalize_word. Нормализованные слова, не являющиеся стоп-словами, сохраняются в списке и возвращаются; search_sentences (query) ищет предложения во фрейме данных, содержащие все слова из заданного запроса (query). Для каждой строки во фрейме данных проверяется, содержатся ли все слова из запроса в нормализованном тексте. Если совпадения найдены, предложение добавляется в список found_sentences. Если совпадения не найдены, возвращается сообщение об ошибке.

При реализации было решено использовать библиотеки NLTK и PyMorphy2 для высокой точности и надежности различных видов анализа текста, поскольку в медицинских рекомендациях и пособиях редко требуется извлекать и анализировать именованные сущности, какие возможности предоставляет библиотека Natasha [18, 19].

Для создания рекомендательной системы была выбрана среда разработки VS Code ввиду ее удобного интерфейса и широкой поддержки расширений, обеспечивающих разнообразные инструменты для работы с Python и его библиотеками, включая отладку, автодополнение кода и управление зависимостями.

Для реализации также был выбран фреймворк Django, который предоставляет готовые компоненты, такие как аутентификация, маршрутизация и работа с базами данных, упрощая процесс разработки. Для базы данных (рис. 3) было решено выбрать MySQL, поскольку она легко интегрируется с интерфейсом API, обладает простотой использования, гибкостью и высокой производительностью.

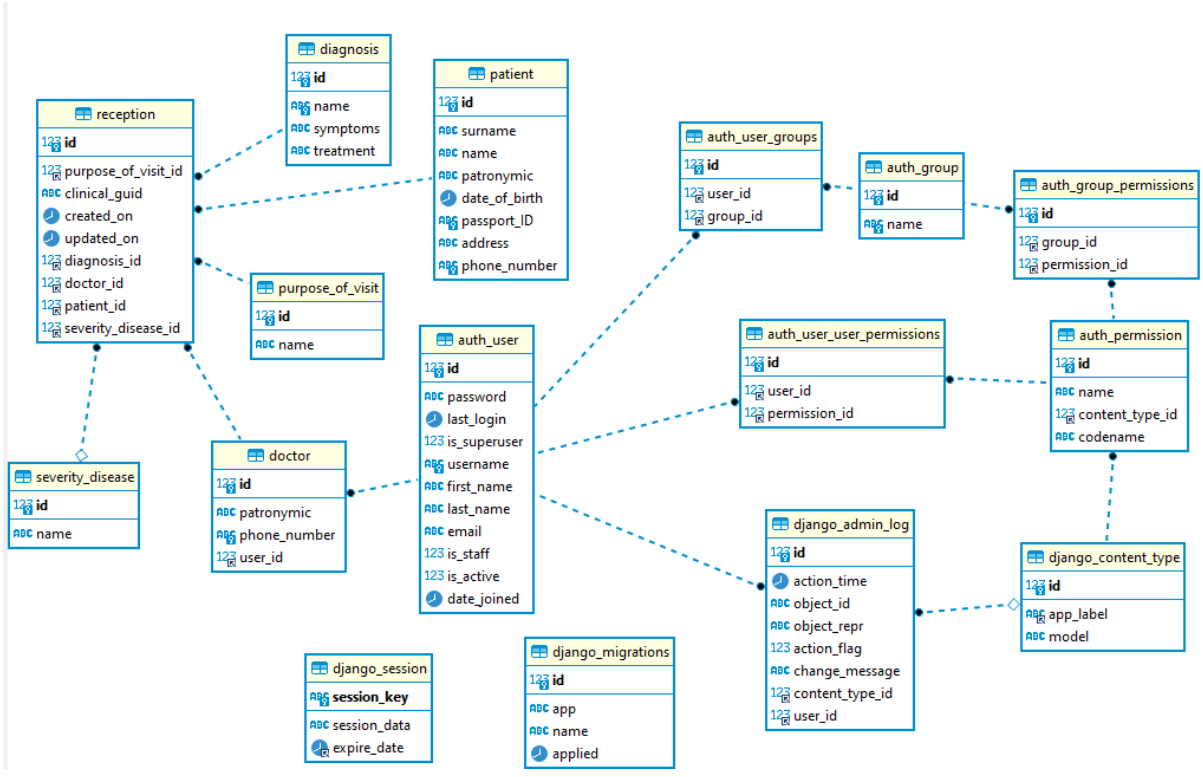

Рис. 3 Схема базы данных.

Результаты вычислительных экспериментов и интерпретация результатов

В ходе эксперимента проведены оценка временных затрат и релевантность выдаваемых рекомендаций с использованием трех реализованных подходов, заявленных ранее (табл. 1).

Таблица 1

Результаты вычислительного эксперимента

Подход	Временные затраты	Релевантность
Поиск по регулярным выражениям	Ожидание выдачи ответа и генерации текста занимает до 3 сек.	В качестве результата выводится текст по нужной тематике, содержащий введенные слова в различных формах и последовательностях
Генеративный подход	Ожидание выдачи ответа и генерации текста занимает до 2 мин.	В качестве результата выводится несвязный текст с некорректной грамматикой
Интеллектуальный поиск	Ожидание выдачи ответа и генерации текста занимает до 5 сек.	В качестве результата выводится текст, содержащий введенное слово и близкие по контексту к нему термины

Для генеративного подхода требуется большой объем данных для обучения, поэтому в данном эксперименте метод показал неэффективный результат (рис. 4).

```
prompt = 'температура'
length = 500
recommendation = generate_text(model, tokenizer, prompt, length)
print(recommendation)
```

```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Pl
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
температурату Эдытра пладовео защеступии: утлия просмер в дазаботчееск задь инік. Наеде бребы роспуй вбрениета.
```

Рис. 4 Применение генеративного подхода при поиске рекомендаций.

Интеллектуальный поиск требует обработки больших текстовых наборов данных и позволяет выделить контекстуальные связи между словами, однако для существующего корпуса предложений он дал релевантные ответы. Для более точного поиска по тексту рекомендуется использовать сочетание методов интеллектуального поиска и поиска по регулярным выражениями.

В русском языке существуют различные группы родственных слов, имеющие разные формы и контекстные значения. Тезаурусы не лишены ошибок, наиболее распространенные из них включают пропущенные или неправильные связи между словами, устаревшие значения, отсутствие новых или разговорных выражений. Векторные представления слов позволяют извлекать новые знания из текстовых коллекций, открывая перспективы для улучшения автоматической обработки текстов на естественном языке [20].

Сравнительный анализ временных затрат на поиск клинических рекомендаций в документе вручную, с помощью навигации и с помощью предложенного алгоритма показал преимущество предложенного решения (табл. 2).

По результатам эксперимента можно сделать заключение, что с применением автоматизированного поиска с учетом индивидуальных особенностей существенно сократилось время, затрачиваемое врачами на поиск справочной информации о заболеваниях или клинических рекомендаций, связанных с лечением пациента. Эти факторы способствуют повышению эффективности работы врачей.

Таблица 2

Сравнение временных затрат на поиск рекомендаций вручную, с помощью навигации и алгоритма

Способ	Временные затраты
Вручную	В среднем человек может просмотреть и оценить содержимое каждой страницы примерно за 10–15 сек. Для документа со 100 страницами потребуется примерно до 16–25 мин.
Навигация в программах по просмотру документа	Для опытного пользователя встроенной навигации время поиска информации в документе со 100 страницами может составлять до 5–10 мин. Однако поиск осуществляется по определенной последовательности ключевых слов или символов, а совместно встречающиеся слова в одном предложении могут не учитываться, если они расположены не друг за другом и имеют иные формы
Разработанный алгоритм	Время поиска информации клинических рекомендаций в документе со 100 страницами может составлять до 1–2 мин.

Заключение

Для совершенствования процесса поиска рекомендаций при лечении пациентов с заболеваниями органов дыхания была разработана рекомендательная система.

В ходе исследования были рассмотрены различные подходы к рекомендательным системам в медицине, их классификация, методы диагностики и лечения, анализировались клинические рекомендации в этой области. В области бронхолегочных заболеваний был проведен аналитический обзор, включающий классификацию, методы диагностики и лечения, анализ клинических рекомендаций. Полученные знания послужили основой для создания механизма генерации рекомендаций и разработки рекомендательной системы.

В результате анализа были выявлены ограничения в выборе решений для предобработки и генерации медицинских клинических рекомендаций, а также зависимость от иностранных программных продуктов.

Для поиска и отбора медицинских рекомендаций была выбрана модель дистрибутивной семантики Word2vec как наиболее рациональная. Для этой модели были разработаны алгоритм, который сочетал поиск совпадающих предложений по регулярным выражениям, и модель поиска дистрибутивно-семантических связей Word2vec. Такой подход позволил упростить и ускорить поиск информации о лечении и профилактике бронхолегочных заболеваний для специалистов медицинской отрасли. Это, в свою очередь, положительно сказалось на решении проблемы ограниченного времени приема для врачей.

Авторами работы был создан алгоритм, который генерирует индивидуальные рекомендации для пациентов. Для этого использовался алгоритм поиска совпадающих предложений по регулярным выражениям.

Также были разработаны архитектура и структура базы данных для хранения информации о пациентах, записях приема и диагнозах. Для удобства использования системы врачами и медицинским персоналом было разработано веб-приложение с интуитивным интерфейсом. Реализация проекта осуществлялась с использованием следующих технологий: IDE VS Code, DBeaver для работы с MySQL, язык программирования Python 3.9, фреймворк Django для создания веб-приложения, инструментарий Bootstrap для создания GUI, а также различные библиотеки, такие как datetime, pyperclip, nltk, os, pandas, pymorphy2, для обработки данных и анализа текста.

Для оценки качества разработанного ПО были проведены тестирование на основе экспериментальных данных и анализ результатов по ГОСТ 28195-89. Анализ эффективности показал, что применение автоматизированного поиска с учетом индивидуальных особенностей

значительно сокращает время, затрачиваемое врачами на поиск справочной информации о бронхолегочных заболеваниях или клинических рекомендациях по лечению пациента.

Благодарности

Исследование выполнено в рамках проекта при поддержке гранта РНФ 22-19-00471. Авторы считают нужным отметить работы коллег по сходной тематике [21–30], оказавшие положительное влияние на данное исследование.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

[1] Грибова В. В., Кульчин Ю. Н., Петряева М. В., Окунь Д.Б., Ковалев Р. И., Шалфеева Е. А. Интеллектуальная система поддержки принятия врачебных решений по дифференциальной диагностике и лечению Covid-19 // Вестник РАН. 2022. Т. 92. № 8. С. 781–789. [[V. V. Gribova, Yu. N. Kulchin, M. V. Petryaeva, D. B. Okun, R. I. Kovalev, E. A. Shalfeeva, “Intelligent system for supporting medical decisions on differential diagnostics and treatment of Covid-19,” (in Russian) // Vestnik RAN, vol. 92, no. 8, pp. 781-789, 2022.]].

[2] Гришина Л. С., Болодурина И. П. Разработка модели генерации клинических рекомендаций для пациентов на основе неструктурированных текстовых данных // Научно-технический вестник Поволжья. 2023. № 8. С. 53–56. [[L. S. Grishina, I. P. Bolodurina, “Development of a model for generating clinical recommendations for patients based on unstructured text data,” (in Russian) // Scientific and Technical Volga region Bulletin, no. 8, pp. 53-56, 2023.]].

[3] Цурко В. В. Рекомендательные системы в здравоохранении // Управление большими системами. 2019. С. 61–73. [[V. V. Tsurko “Recommender systems in healthcare,” (in Russian) // Large-Scale Systems Control, pp. 61-73, 2019.]].

[4] Кобринский Б. А. Интеллектуальные рекомендательные системы для медицины: особенности и ограничения // Искусственный интеллект и принятие решений. 2022. № 3. С. 51–62. [[B. A. Kobrinsky “Intelligent recommender systems for medicine: features and limitations,” (in Russian) // Artificial Intelligence and Decision Making, no. 3, pp. 51-62, 2022.]].

[5] Гусев А. В, Зарубина Т. В. Поддержка принятия врачебных решений в медицинских информационных системах медицинской организации// Врач и информационные технологии. 2017. № 2. С. 60–72. [[A. V. Gusev, T. V. Zarubina “Support for medical decision-making in medical information systems of a medical organization,” (in Russian) // Medical Doctor and Information Technologies, no. 2, pp. 60-72, 2017.]].

[6] Thi Ngoc Trang Tran, Alexander Felfernig, Christoph Trattner, Andreas Holzinger, “Recommender systems in the healthcare domain: state-of-the-art and research issues” // Journal of Intelligent Information Systems. 2021. Vol. 57. No. 8. Pp. 1–31. DOI: 10.1007/s10844-020-00633-6.

[7] Stark B., Knahl C., Aydin M., Elish K. A literature review on medicine recommender systems // International Journal of Advanced Computer Science and Applications. 2019. Vol. 10. No. 8. Pp. 6–13. DOI: 10.14569/IJACSA.2019.0100802.

[8] Kamyshev K. V., Kureichik V. M., Borodyanskiy I. M. Review of the recommender systems application in cardiology // Cardiometry. 2020. No. 16. Pp. 97–105. DOI: 10.12710/cardiometry.2020.16.97105.

[9] Ochoa J. G. D., Csiszár O., Schimper Th. Medical recommender systems based on continuous-valued logic and multi-criteria decision operators, using interpretable neural networks // BMC Medical Informatics and Decision Making. 2021. Vol. 21. No. 1. DOI: 10.1186/s12911-021-01553-3.

[10] Козлов С. В., Светлаков А. В. Применение регулярных выражений для обработки текстовых данных // International Journal of Open Information Technologies. 2022. Т. 10. № 9. С. 82–89. [[S.V. Kozlov, A. V. Svetlakov “Using regular expressions to process text,” (in Russian) // International Journal of Open Information Technologies, vol. 10, no. 9, pp. 82-89, 2022.]]

[11] Satav M. S., Varade T., Kothavale Dh., Thombare S., Lokhande P. Data extraction from invoices using computer vision // Proc. IEEE 15th International Conference on Industrial and Information Systems (ICI-IS). 2020. Pp. 316–320.

[12] Rajput Sh., Mehta N., Singh A., Raghunandan H. Keshavan, Vu T., Heldt L., Hong L., Tay Y., Vinh Q. Tran, Samost J., Kula M., Ed H. Chi, Sathiamoorthy M. Recommender systems with generative retrieval // Proc. 37th Conference on Neural Information Processing Systems (NeurIPS). 2023. URL: https://arxiv.org/pdf/2305.05065.

[13] Шахмаметова Г. Р., Зулкарнеев Р. Х., Евграфов А. А. Методы обработки текстовых данных в системе принятия клинических решений при диагностике болезней органов дыхания // Информационные технологии интеллектуальной поддержки принятия решений (ITIDS'2019): Труды VII Всеросс. научн. конф.: В 3 т. Т. 2. Уфа : УГАТУ, 2019. С. 245–248. EDN LMAEVT. [[G.R. Shakhmametova, R. Kh.Zulkarneev, A. A. Evgrafov, (in Russian), // Proc. Information Technologies for Intelligent Decision Support (ITIDS'2019), Ufa, Russia, 2019, pp. 245-248. EDN LMAEVT.]].

[14] Власов Д. Ю., Пальчунов Д. Е., Степанов П. А. Автоматизация извлечения отношений между понятиями из текстов естественного языка // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2010. Т. 8. № 3. С. 23–33. [[D. Yu. Vlasov, D. E. Palchunov, P. A. Stepanov “Automation of extraction of relations between concepts from natural language texts,” (in Russian) // Vestnik NSU. Series: Information Technologies, vol. 8, no. 3, pp. 23-33, 2010.]].

[15] Жеребцова Ю. А., Чижик А. В. Сравнение моделей векторного представления текстов в задаче создания чатбота // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2020. Т. 18. № 3. С. 16–34. [[Yu. A. Zherebtsova, A. V Chizhik. “Text vectorization methods for retrieval-based chatbot,” (in Russian) // NSU Vestnik. Series: Linguistics and Intercultural Communication, vol. 18, no. 3, pp. 16-34, 2020, . DOI 10.25205/1818-7935-2020-18-3-16-34.]].

[16] Тихомиров М. М. Методы автоматизированного пополнения графов знаний на основе векторных представлений: дис…. канд. физ.-мат. наук: 05.13.11. М., 2022. 119 с. [[Tikhomirov M. M. Methods of Automated Replenishment of Graph Knowledge Based on Vector Representations (in Russian): Diss. Cand. of Phys. and Math. Sci.: 05.13.11. Moscow, 2022.]].

[17] Baroni M., Lenci A. How we BLESSed distributional semantic evaluation // Proc. GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, Association for Computational Linguistics. 2011. Pp. 1–10.

[18] Жаббарова Р. У., Бурнашев Р. Ф. Инструментарий обработки лингвистической информации // Science and Education. 2023. Т. 4. № 4. С. 654–664. [[R. U. Zhabbarova, R. F. Burnashev “Tools for processing linguistic information,” (in Russian) // Science and Education, vol. 4, no. 4, pp. 654–664, 2023.]].

[19] Шульман В. Д., Максименко О. Е., Волхонцева П. Д., Анализ программных средств морфологического анализа // Международный журнал гуманитарных и естественных наук. 2022. Т. 3–2. № 66. С. 166–170. [[V. D. Shulman, O. E. Maksimenko, P. D. Volkhontseva “Analysis of morphological analysis soft-ware tools,” (in Russian) // International Journal of Humanities and Natural Sciences, vol. 3-2, no. 66, pp. 166-170, 2022.]].

[20] Loukachevitch N. V. Corpus-based check-up for thesaurus // Proc. I57th Annual Meeting of the Association for Computational Linguistics, 2019. Pp. 5773–5779.

[21] Зиновьев М. С., Нургаянова О. С. Прогнозирование вероятности развития диабетической ретинопатии у пациентов с сахарным диабетом: анализ методов машинного обучения // СИИТ. 2024. Т. 6. № 3(18). С. 95–101. EDN VLFFLP. [[Zinoviev M. S., Nurgayanova O. S. “Predicting the probability of developing diabetic retinopathy in patients with diabetes mellitus: analysis of machine learning methods” // SIIT. 2024. Vol. 6, No. 3(18), pp. 95-101. EDN VLFFLP. (In Russian).]]

[22] Шапошникова А. С., Богданов М. Р. Определение сердечного ритма плода по неинвазивному ЭКГ с применением различных фильтров // СИИТ. 2023. Т. 5. № 6(15). С. 32-37. EDN WBBOVK. [[Shaposhnikova A. S., Bogdanov M. R. “Determination of fetal heart rate by non-invasive ECG” // SIIT. 2023. Vol. 5, No. 6(15), pp. 32-37. EDN WBBOVK. (In Russian).]]

[23] Зиновьев М. С., Нургаянова О. С. Оценка индивидуального риска развития сахарного диабета второго типа и возможных осложнений // СИИТ. 2023. Т. 5. № 4(13). С. 101–110. EDN HIIXFH. [[Zinoviev M. S., Nurgayanova O. S. “Assessment of individual risk of developing type 2 diabetes mellitus and possible complications” // SIIT. 2023. Vol. 5, No. 4(13), pp. 101-110. EDN HIIXFH. (In Russian).]]

[24] Шалфеева Е. А. Методология производства жизнеспособных систем доверительного искусственного интеллекта // СИИТ. 2023. Т. 5. № 4(13). С. 28–49. EDN CJTKQH. [[Shalfeeva E. A. “Methodology for the production of viable systems of trustworthy artificial intelligence” // SIIT. 2023. Vol. 5, No. 4(13), pp. 28-49. EDN CJTKQH. (In Russian).]]

[25] Юсупова Н. И., Нургаянова О. С., Зулкарнеев Р. Х. Формализация этапов риск-анализа в СППР с учетом оценок клинических рисков при бронхолегочных заболеваниях // СИИТ. 2023. Т. 5. № 1(10). С. 11–24. EDN KHIIHT. [[Yusupova N. I., Nurgayanova O. S., Zulkarneev R. Kh. “Formalization of risk analysis stages in decision support system taking into account clinical risk assessments for bronchopulmonary diseases” // SIIT. 2023. Vol. 5, No. 1(10), pp. 11-24. EDN KHIIHT. (In Russian).]]

[26] Шахмаметова Г. Р., Христодуло А. Д., Береговая С. П. Анализ эндокринологических данных на основе моделей классификации // СИИТ. 2022. Т. 4. № 2(9). С. 30–36. EDN LBZVZL. [[Shakhmametova G. R., Khristodullo A. D., Beregovaya S. P. “Analysis of endocrinological data based on classification” // SIIT. 2022. Vol. 4, No. 2(9), pp. 30-36. EDN LBZVZL. (In Russian).]]

[27] Насыров Р. В. Причинный подход к построению бионических вычислений на основе рекурсивных моделей анализа данных // СИИТ. 2022. Т. 4. № 1(8). С. 27–36. EDN UOMMOU. [[Nasyrov R. V. “Causal approach to the construction of bionic computations based on recursive models” // SIIT. 2022. Vol. 4, No. 1(8), pp. 27-36. EDN UOMMOU. (In Russian).]]

[28] Слепов Д. С. Анализ тенденций развития информационных технологий в условиях распространения новой коронавирусной инфекции (на примере Республики Башкортостан) // СИИТ. 2021. Т. 3. № 3(7). С. 30–36. EDN QOFNKC. [[Slepov D.S. “Analysis of trends in the development of information technologies in the context of the spread of a new coronavirus infection (on the example of the Republic of Bashkortostan)” // SIIT. 2021. Vol. 3, No. 3(7), pp. 30-36. EDN QOFNKC. (In Russian).]]

[29] Николаева М. А., Агадуллина А. И. Математическое обеспечение системы анализа гериатрических рисков // СИИТ. 2020. Т. 2. № 2(4). С. 66–72. EDN NODUDY. [[Nikolaeva M. A., Agadullina A. I. “Mathematical support for the geriatric risk analysis system” // SIIT. 2020. Vol. 2, No. 2(4), pp. 66-72. EDN NODUDY. (In Russian).]]

[30] Бухарбаева Л. Я., Франц М. В., Кондрова Н. С. Информационные технологии оценки бремени болезней и формирования оптимальных профилактических программ // СИИТ. 2020. Т. 2. № 1(3). С. 67–72. EDN UFNOJX. [[Bukharbaeva L. Ya., Franz M. V., Kondrova N. S. “Information technologies for assessing the burden of diseases and forming optimal preventive programs” // SIIT. 2020. Vol. 2, No. 1(3), pp. 67-72. EDN UFNOJX. (In Russian).]]

Поступила в редакцию 22 октября 2024 г.

МЕТАДАННЫЕ / METADATA

Title: Medical recommendation system based on automatic knowledge extraction from texts.

Abstract: The limited time for doctors' appointments, requiring real-time decision-making, is accompanied by regular changes in clinical guidelines based on new research. Therefore, the relevance of the problem situation is due to the need for information support for a health worker in order to reduce the time it takes to search for recommendations for diagnosing and prescribing treatment for patients with various diseases. Modern development of text analysis tools allows automating the process of searching for recommendations. The article proposes an approach to building a recommender system that can effectively search for information on the correct diagnosis, treatment and prevention of diseases. In the course of the study, an analysis of the current state of the problem situation was carried out, in particular, approaches to creating recommended systems in medicine. For searching

and selecting medical recommendations, the Word2vec distribution semantics model was chosen as the most rational one, an algorithm was developed in the form of a combination of searching for matching sentences by regular expressions and the Word2vec model for searching for distribution-semantic relationships. Limitations in choosing solutions for pre-processing and generating medical clinical recommendations were identified, an algorithm was developed that can generate individual recommendations for patients based on the analysis of matching sentences by regular expressions. The authors of the article proposed the architecture and structure of a database for storing information about patients, appointment records and diagnoses. It was demonstrated that the created recommendation system with an integrated Word2vec distribution semantics algorithm reduces the time spent by medical personnel on searching for reference information and can be useful for specialists in the medical industry. The proposed solutions have practical value and can serve as a basis for further research and development of automated systems in medicine.

Key words: medical recommender systems, machine learning, regular grammars, natural language processing, methods of text data analysis, distributional semantics, bronchopulmonary diseases

Язык статьи / Language: Русский / Russian.

Поддержка/Support: РНФ, грант 22-19-00471.

Об авторах / About the authors:

КОРОВИН Евгений Алексеевич
ВШЭ Московский институт электроники и математики им. А. Н. Тихонова (ВШЭ МИЭМ), Россия.
Магистрант. Дипл. бакалавр по программной инженерии (Уфимск. ун-т науки и технол., 2023). Иссл. в обл. управления, сист. анализа, интел. обработки данных.
E-mail: arkvinst@gmail.com

ЧИГЛИНЦЕВА Светлана Андреевна
ВШЭ Московский институт электроники и математики им. А. Н. Тихонова (ВШЭ МИЭМ), Россия.
Магистрант. Дипл. бакалавр по программной инженерии (Уфимск. ун-т науки и технол., 2023). Иссл. в обл. управления, сист. анализа, интел. обработки данных.
E-mail: s_chiglintseva@invox.ru

САЗОНОВА Екатерина Юрьевна
Уфимский университет науки и технологий, Россия.
Доц. каф. вычислительной математики и кибернетики. Дипл. экон.-математик (Уфимск. гос. авиац. техн. ун-т, 2011). Канд. техн. наук по системному анализу, управлению и обработки информации (там же, 2015). Иссл. в обл. управления, сист. анализа, интел. обр. данных.
E-mail: sazonova.eyu@ugatu.su
ORCID: http://orcid.org/0000-0001-8834-2635
URL: https://elibrary.ru/author_profile.asp?authorid=108068

СМЕТАНИНА Ольга Николаевна
Уфимский университет науки и технологий, Россия.
Проф. каф. вычислительной математики и кибернетики. Дипл. инж.-электрик (Уфимск. авиац. ин-т, 1985). Д-р техн. наук по управлению в соц. и экон. системах (Уфимск. гос. авиац. техн. ун-т, 2012). Иссл. в обл. интелл. поддержки принятия решений.
E-mail: smetanina.on@ugatu.su
ORCID: http://orcid.org/0000- 0001-6970-1110
URL: https://elibrary.ru/author_profile.asp?authorid=161318

KOROVIN Evgeniy Alekseevich
HSE Tikhonov Moscow Institute of Electronics and Mathematics, Russia.
Master's student. Bachelor's degree in software engineering (Ufa Univ. of Science & Technology, 2023). Research in the field of management, systems analysis, intelligent data processing.
E-mail: arkvinst@gmail.com

CHIGLINTSEVA Svetlana Andreevna
HSE Tikhonov Moscow Institute of Electronics and Mathematics, Russia.
Master's student. Bachelor's degree in software engineering (Ufa Univ. of Science & Technology, 2023). Research in the field of management, systems analysis, intelligent data processing.
E-mail: s_chiglintseva@invox.ru

SAZONOVA Ekaterina Yurevna
Ufa University of Science & Technologies, Russia.
Assoc. Prof. of Computational Mathematics and Cybernetics dept. Dipl. Economist-Mathematician (Ufa State Aviation Tech. Univ., 2011). Cand. Tech. Sci. on System Analysis, Management and Information Processing (ibid., 2015). Research in the field of management, systems analysis, data mining.
E-mail: sazonova.eyu@ugatu.su
ORCID: http://orcid.org/0000-0001-8834-2635
URL: https://elibrary.ru/author_profile.asp?authorid=108068

SMETANINA Olga Nikolavna
Ufa University of Science & Technologies, Russia.
Prof. Computational Mathematics and Cybernetics dept. Dipl. Eng.-Electrician (Ufa Aviat. Institut, 1985). Dr. Technical Sciences in social and economic management systems (Ufa State Aviat. Tech. University, 2012). Research in the field of intelligent decision support.
E-mail: smetanina.on@ugatu.su
ORCID: http://orcid.org/0000- 0001-6970-1110
URL: https://elibrary.ru/author_profile.asp?authorid=161318