Разработка метода структурирования данных и знаний клинических рекомендаций

Г.Р. Шахмаметова
Факультет информатики и робототехники
Уфимский государственный авиационный технический университет
Уфа, Россия
e-mail: shakhgouzel@mail.ru

Е.В. Худоба
Факультет информатики и робототехники
Уфимский государственный авиационный технический университет
Уфа, Россия
e-mail: eugchud@gmail.com

Аннотация[1]

В статье проведен обзор существующих программных решений в области анализа неструктурированных текстов (как неспециализированных, так и медицинского характера). По результатам обзора выделены недостатки существующих разработок и выявлена потребность в новом программном решении для анализа клинических рекомендаций. Поставлена задача разработки данного решения, определены его ключевые характеристики, описан метод структуризации данных и знаний клинических рекомендаций.

1. Введение

В течение последних десятилетий практически во всех областях человеческой деятельности (научных исследованиях, экономике, предпринимательской деятельности) многократно возросли объемы хранимой, обрабатываемой и передаваемой информации. Это привело к существенному росту интереса исследователей к методам и алгоритмам обработки данных и знаний.

По степени организованности данные можно условно разделить на две категории:

- Структурированные, примерами которых являются базы данных, логи информационных систем, данные датчиков и сенсоров.
- Неструктурированные, в качестве примеров которых можно привести текстовые данные, изображения, видео.

По оценкам специалистов [1], около 80–90% всей используемой в организациях информации представлено в неструктурированной форме. В связи с этим возникает потребность в сокращении затрат трудовых, временных, финансовых и других видов ресурсов, необходимых для обработки такой информации. Наиболее эффективным способом достижения этой цели является приведение такой информации к структурированной форме (структуризация данных).

Еще в середине XX века среди исследователей (в частности, Х. П. Лун и др.) возник интерес к способам извлечения и классификации данных в неструктурированных текстах, однако лишь в последние несколько десятилетий появились необходимые для таких исследований технологии [2].

Значительное повышение интереса к этой области исследований было вызвано появлением таких направлений, как Big Data, Data Mining и Natural Language Processing.

Одной из наиболее важных сфер, в которых возможно применение технологий структуризации данных, представленных в текстовой форме, является здравоохранение. В частности, большое практическое значение в контексте усовершенствования услуг системы здравоохранения имеет задача анализа клинической документации, т. е. таких документов, как медицинские карты, результаты обследований, журналы оперативных вмешательств и т. д.

Среди наименее исследованных на сегодняшний день задач в данной области можно назвать задачу анализа текстов клинических рекомендаций.

Клинические рекомендации представляют собой специализированные документы, разрабатываемые с целью поддержки принятия решения практикующим врачом для обеспечения надлежащей медицинской помощи в конкретной клинической ситуации. Фактически данный документ является руководством специалиста по ведению пациента, его диагностике и лечению.

Клинические рекомендации содержат неструктурированные данные и знания, которыми руководствуется специалист при назначении лечения, обследований и принятии других решений, влияющих на исход заболевания пациента. В исходном виде эти данные и знания непригодны для

Труды Седьмой всероссийской научной конференции "Информационные технологии интеллектуальной поддержки принятия решений", 28-30 мая, Уфа – Ставрополь - Ханты-Мансийск, Россия, 2019

автоматизированной обработки, в связи с чем анализ клинических рекомендаций в медицинской практике производится вручную. Если возможно привести эти данные и знания к структурированному виду, то их можно было бы использовать в системах поддержки принятия клинических решений (СППКР) при диагностике и выборе траектории лечения заболеваний.

2. Анализ методов и средств извлечения данных и знаний из неструктурированных данных

В области структуризации текстовых данных как для исследовательских целей, так и для решения прикладных задач было разработано большое количество программных решений.

2.1. Решения для анализа текстов общего содержания

- SAS Text Miner — интегрированный компонент системы SAS, созданный для анализа текстовых данных, предоставляет большой набор инструментов лингвистического и аналитического моделирования, разработанных специально для обнаружения и извлечения знаний из коллекций текстовой информации [3].
- GATE (General Architecture for Text Engineering) — система обработки естественного языка с открытым исходным кодом, использующая наборы компонентов на языке Java [4].
- STATISTICA Text Miner – дополнительное расширение STATISTICA Data Miner, предназначенное для извлечения знаний из неструктурированных текстов [5].
- Natural Language Toolkit (NLTK) — пакет библиотек и программ для символьной и статистической обработки естественного языка, написанных на языке программирования Python, содержащий графические представления и примеры данных [6].

2.2. Решения для анализа текстов медицинской тематики

Рассмотренные выше программные решения ориентированы на обработку текстов общего характера, например, таких как новостные сообщения. В то же время следует отметить, что стилистика клинических текстов сильно отличается от стилистики текстов других предметных областей, поэтому для их анализа требуется значительная доработка существующих методов и инструментов по анализу текстов на естественном языке. В связи с этим анализ клинических текстов был выделен в отдельное направление исследований.

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

- UMLS (Unified Medical Language System) — средство для разработки компьютерных систем анализа биомедицинской информации и других видов информации в сфере здравоохранения. Разработана в 1986 году в Национальной медицинской библиотеке США (National Library of Medicine, NLM) [7].
- MedLEE (Medical Language Extraction and Encoding System) — система, осуществляющая извлечение, структурирование и кодирование клинической информации, содержащейся в различных видах медицинских отчетов (напр., по рентгеновским, маммографическим и эхокардиологическим исследованиям) [8].
- cTAKES (clinical Text Analysis and Knowledge Extraction System) — система обработки естественных языков с открытым исходным кодом, осуществляющая извлечение клинической информации из неструктурированных текстов электронных медицинских карт [9].

Все из вышеперечисленных систем обладают существенным недостатком в рамках нашей задачи: ни одна из них не обладает встроенной поддержкой русского языка.

2.3. Решения автоматического построения онтологий на основе текстовых документов

Возможным средством решения задачи для представления данных и знаний клинических рекомендаций является онтология - описание предметной области, представленное в виде концептуальной схемы. Мы рассмотрели наиболее известные средства автоматической генерации онтологий на основе текстовых файлов.

- Text-To-Onto — программное решение, разработанное исследователями University of Karlsruhe, которое осуществляет автоматическое построение онтологий на основе текстов на естественном языке путем выявления в них ключевых понятий и обнаружения связей между ними [10].
- DOG4DAG (Dresden Ontology Generator for Directed Acyclic Graphs) – средство автоматической генерации онтологий на основе текстов на естественном языке. Оно представлено в форме плагина для Protégé 4.1 и OBOEdit 2.1. Данный плагин позволяет использовать в качестве входных данных статьи PubMed, Web-страницы или PDF-документы. Генерация онтологии в DOG4DAG осуществляется при помощи построения иерархической модели классов, связанных отношениями вида “is subclass of” [11].

В результате анализа средств автоматической генерации онтологий (таких как, напр., ASIUM, Syndikate, WebKB и др.) было выявлено, что поддержка подавляющего большинства из них в настоящее время прекращена, многие из них недоступны для загрузки, а также ни один из них не поддерживает обработку текстов на русском языке.

В связи с этим для решения проблемы структуризации данных и знаний клинических рекомендаций необходима разработка алгоритма извлечения данных и знаний из русскоязычных текстов медицинской тематики.

3. Постановка задачи

Проведенный анализ современного состояния исследований показал отсутствие готовых решений в этой области. Поэтому необходима разработка метода и алгоритма структуризации данных и знаний русскоязычных текстов клинических рекомендаций и приведения их к виду, пригодному для дальнейшей обработки в системах поддержки принятия клинических решений, а также их программная реализация.

Разрабатываемое программное решение должно обладать следующими характеристиками:

- программное решение должно принимать на вход тексты клинических рекомендаций на русском языке;
- результатом обработки текста должен являться набор правил, пригодных к использованию в системах поддержки принятия клинических решений.

4. Разработка метода структуризации данных и знаний клинических рекомендаций

4.1. Общая структура метода

Для процесса обработки текстов клинических рекомендаций выделены следующие основные этапы:

- определение ключевых слов (осуществляемое автоматически при помощи специализированных алгоритмов либо вручную);
- построение карты понятий на основе ключевых слов;
- выделение в карте понятий правил и их представление в виде, пригодном для использования в СППКР.

Диаграмма, иллюстрирующая перечисленные выше этапы, представлена на рис. 1.

Рис. 1. Схема процесса структуризации данных и знаний клинических рекомендаций

4.2. Карта понятий

Карта понятий представляет собой ориентированный граф, в вершинах которого записываются понятия предметной области, а в ребрах – отношения между ними. Отношения между понятиями могут быть как таксономическими (т. е. формирующими иерархию понятий), так и других видов.

Ключевой особенностью карт понятий является возможность отображения правил связей объектов, в основе которых лежат условные конструкции. Их отображение на диаграмме производится при помощи пунктирных линий; подпись связей условной части правила сопровождается префиксом «У», а заключения правила – префиксом «З».

Пример карты понятий приведен на рисунке 2.

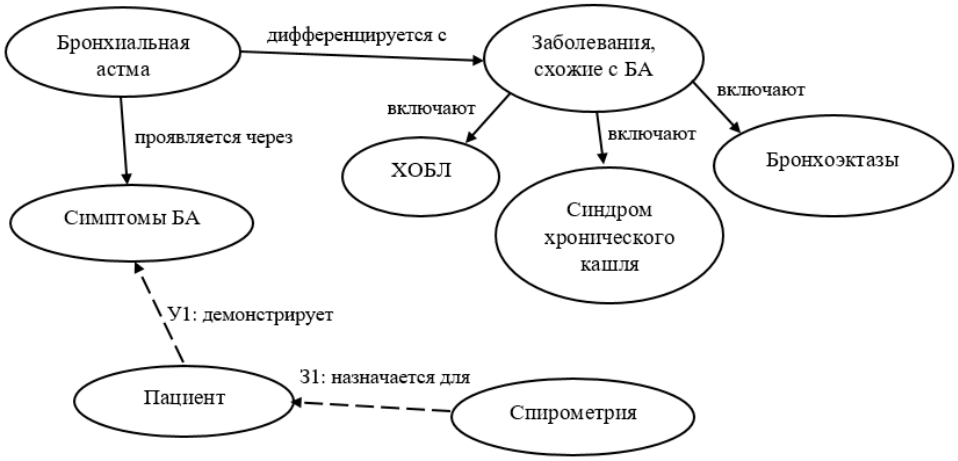

Рис. 2. Пример карты понятий

На основе карты понятий с учетом условных связей формируются продукционные правила вида «Если… то…», пригодные далее для использования в системах поддержки принятия клинических решений.

5. Заключение

Проведенный анализ существующих решений для задачи автоматического извлечения данных и знаний из текстов клинических рекомендаций показал, что ни одно из имеющихся в настоящее время программных средств не подходит для решения рассматриваемой задачи. В связи с этим был сделан вывод о необходимости разработки нового метода структуризации данных и знаний клинических рекомендаций и реализации его в виде программного решения.

Предложенный метод отличает использование нового средства представления данных и знаний в структурированном виде, названного «карта понятий»; данное средство позволяет представлять связи между понятиями, содержащие условную и заключительную часть. Так как одним из ключевых этапов разрабатываемого метода автоматического извлечения данных и знаний из клинических рекомендаций является генерация правил из полученных карт понятий, то возможно дальнейшее применение этих правил в базах знаний систем поддержки принятия клинических решений.

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00780, 19-07-00709.

Список используемых источников

1. Shilakes, Christopher C.; Tylman, Julie (16 Nov 1998). "Enterprise Information Portals" (PDF). Merrill Lynch.
2. Grimes, Seth. "A Brief History of Text Analytics". B Eye Network. Retrieved June 24, 2016.
3. Text Mining Software, SAS Text Miner | SAS [Электронный ресурс]. Дата обновления: 18.03.2019. URL: https://www.sas.com/en_us/software/text-miner.html (дата обращения: 18.03.2019).
4. General Architecture for Text Engineering [Электронный ресурс]. Дата обновления: 18.03.2019. URL: https://gate.ac.uk/ (дата обращения: 18.03.2019).
5. STATISTICA Text Miner [Электронный ресурс]. Дата обновления: 18.03.2019. URL: http://statsoft.ru/products/STATISTICA_Data_Miner/STATISTICA_Text_Miner/ (дата обращения: 18.03.2019).
6. Natural Language Toolkit — NLTK 3.4 documentation [Электронный ресурс]. Дата обновления: 18.03.2019. URL: https://www.nltk.org/ (дата обращения: 18.03.2019).
7. Unified Medical Language System (UMLS) [Электронный ресурс]. Дата обновления: 18.03.2019. URL: https://www.nlm.nih.gov/research/umls/ (дата обращения: 18.03.2019).
8. MedLEE | MedLingMap [Электронный ресурс]. Дата обновления: 18.03.2019. URL: http://www.medlingmap.org/taxonomy/term/80 (дата обращения: 18.03.2019).
9. Apache cTAKES - clinical Text Analysis Knowledge Extraction System [Электронный ресурс]. Дата обновления: 18.03.2019. URL: http://ctakes.apache.org/ (дата обращения: 18.03.2019).
10. Maedche, Alexander; Staab, Stephen (July 2000). " The TEXT-TO-ONTO Ontology Learning Environment" (PDF). ResearchGate.
11. Dresden Ontology Generator for Directed Acyclic Graphs (DOG4DAG) [Электронный ресурс]. Дата обновления: 30.04.2019. URL: http://www.biotec.tu-dresden.de/research/schroeder/dog4dag/ (дата обращения: 30.04.2019