УДК 004.421.6

ОБРОБКА СЛАБОСТРУКТУРОВАНИХ ТЕКСТОВИХ ДАНИХ КЛІНІЧНИХ ПРОТОКОЛІВ ДЛЯ СИСТЕМИ ПІДТРИМКИ ПРИЙНЯТТЯ МЕДИЧНИХ РІШЕНЬ

Бичко Дмитро, Шендрик Віра, Парфененко Юлія

Сумський державний університет

Анотація

Розглянуто підхід до обробки слабоструктурованих текстових даних клінічних протоколів, що зберігаються та розповсюджуються у вигляді файлів у pdf-форматі. У ході роботи вирішено задачу первинної обробки даних клінічних протоколів на прикладі уніфікованого клінічного протоколу первинної, вторинної (спеціалізованої) та третинної (високоспеціалізованої) медичної допомоги. Розроблено метод обробки текстів клінічних протоколів для створення чіткої структури симптомів хвороби та видобування даних при прийнятті медичних рішень.

Abstract

An approach to processing poorly structured clinical protocol text data stored and disseminated as pdf files is considered. In the course of the work, the problem of primary processing of clinical protocol data was solved by the example of a unified clinical protocol of primary, secondary (specialized) and tertiary (highly specialized) medical care. A method for processing the texts of clinical protocols to create a clear structure of the symptoms of the disease and extract data when making medical decisions.

Вступ

На сьогоднішній день існує достатня кількість медичних знань, які накопичені за період існування людства. На ринку інформаційних технологій присутня велика кількість програмного забезпечення, що дозволяє покращити якість медичних послуг. У медичній практиці лікарі використовують клінічні протоколи лікування хвороб, що представлені у вигляді набору слабоструктурованих природомовних текстових даних для встановлення діагнозу та призначення відповідного лікування. Але через відсутність чіткої структури даних цей процес ускладнюється, тому для підвищення швидкості та якості прийняття рішень у ході діагностики та лікування необхідно розробити інформаційну систему підтримки прийняття рішень на основі медичних знань. Для цього необхідно розробити структуру подання даних з уніфікованих клінічних протоколів та реалізувати алгоритми їх обробки. Тому першочерговою задачею є обробка текстів медичних протоколів з метою виділення змістовної інформації та її перетворення у структурований формат, що зробить можливим її автоматичне опрацювання у ході прийняття медичних рішень.

Результати дослідження

У даній роботі розглянуто підхід до обробки слабоструктурованих текстових даних клінічних протоколів, які представлені у вигляді pdf-файлів. Інформація, наведена у клінічних протоколах, розділена на декілька змістовних розділів, але відсутня єдина форма її представлення через різний обсяг та формат інформації, наявність схем та рисунків, що значно ускладнює процес обробки протоколу, перегляду та швидкого пошуку необхідних даних. На сьогодні систематизація даних клінічних протоколів реалізована на рівні каталогізації файлів з повними текстами клінічних протоколів за назвою хвороби чи групи захворювань. Автоматизований пошук необхідних знань виконується за допомогою спрощених пошукових методів (синтаксичний − токенізація, тегування частин мови та семантичний − розпізнавання різних частин мови) аналізу тексту [1], використання специфічного програмного забеспечення [2, 3, 4] або ж вузьконаправлених парсерів, що дозволяють працювати лише з однотипними документами [5]. Наявність даних можливостей дозволяють лише частково задовольняти

потреби користувачів. Проблема видобування, обробки, перетворення та зберігання неструктурованих та слабоструктурованих даних клінічних протоколів для більш швидкої взаємодії користувача з інформаційною системою є досить актуальною. Тому необхідним є структурування даних клінічних протоколів та подання у чітко визначеному форматі, що дозволить швидко оброблювати вхідну інформацію та більш точно діагностувати хворобу, спираючись на вже існуючі експертні знання, що знижує вплив людського фактору на прийняття медичних рішень.

Робота присвячена розробці методу обробки слабоструктурованих текстів уніфікованих клінічних протоколів для перетворення файлів протоколів у таку форму, яка дозволяє їх використання системою підтримки прийняття медичних рішень.

Задача первинної обробки тексту клінічного протоколу потребує його попереднього поділу на чотири базові частини для швидшого конвертування у інші формати. На першому етапі алгоритму методу дані, отримані з файлів pdf-формату, перетворюються на txt за допомогою розробленого парсеру мовою C# з використанням бібліотек iTextSharp.text.pdf та iTextSharp.text.pdf.parser. У результаті отримуємо єдиний txt-файл клінічного протоколу, який містить всю інформацію протоколу у зручному для подальшої обробки форматі. Наступним кроком необхідно обробити цей файл. Для цього проводиться аналіз структури протоколу та виділення основних частин: титульний аркуш (пошук назви хвороби шляхом порівняння назв зі світовою класифікацією МКХ-10), основної частини протоколу (видалення даних до зустрічі з наступним форматом – "цифра один римська або арабська та дві великі літери після них"), видалення даних від літературних джерел і до кінця документу (пошук слова, що починається на "літерат"). У результаті застосування методу первинної обробки тексту клінічного протоколу отримуємо pdf-файл, який містить приблизно у три рази менше інформації, ніж було до його обробки. Він містить у собі один клінічний протокол, в якому залишилася лише змістовна інформація. Наступним є етап екстракції даних з попередньо оброблених протоколів.

У результаті роботи запропоновано метод обробки слабоструктурованих медичних даних на прикладі уніфікованого клінічного протоколу. Виділено інформацію за основними критеріями пошуку та підготовлено файли з медичними протоколами у такій формі, яка дозволить якісно зіставляти дані з медичного протоколу із базою існуючих симптомів для створення структурованого набору даних, який описує хворобу. Сформовано сховище даних оброблених медичних протоколів. У подальших дослідженнях на основі оброблених даних клінічних протоколів, які представлені у вигляді симптомів та їх значень, буде розроблено систему підтримки прийняття медичних рішень.

Список використаних джерел

1. Jensen K, Soguero-Ruiz C, Oyvind MK, Lindsetmo R, Kouskoumvekaki I, Girolami M, et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. Sci Rep 2017 Dec 07;7:46226

2. Patel R, Lloyd T, Jackson R, Ball M, Shetty H, Broadbent M. et al. Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes. BMJ Open 2015; 05 (05) e007504.

3. Wi CI, Sohn S, Rolfes MC, Seabright A, Ryu E, Voge G, et al. Application of a natural language processing algorithm to asthma ascertainment. An automated chart review. Am J Resp Crit Care Med. (2017) 196:430–7. doi: 10.1164/rccm.201610-2006OC

4. Afzal N, Sohn S, Abram S, Scott CG, Chaudhry R, Liu H, et al. Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. J Vasc Surg 2017 Dec;65(6):1753-1761

5. Kung R, Ma A, Dever JB, et al. Mo1043 a natural language processing algorithm for identification of patients with cirrhosis from electronic medical records. Gastroenterology. 2015;148:1071–1072.