Mining Healthcare Web Sites

Praveen Suthar
Arizona State University
psuthar4@asu.edu

Aditya Samant
Arizona State University
amsaman1@asu.edu

Gaurav Kulkarni
Arizona State University
gkulkar5@asu.edu

Kshitiz Lakhotia
Arizona State University
klakhoti@asu.edu

Pulkit Singh Singaria
Arizona State University
psingari@asu.edu

Venkata Sai Manoj Pogadadanda
Arizona State University
vpogadad@asu.edu

***Abstract*—Healthcare is a topic that is important to every person and it is also one where finding relevant and reliable information is crucial for patients. There is an abundance of websites and discussion forums on the internet where information related to various diseases and health issues is available. However, this information is not readily accessible to patients since they would need to parse through a lot of irrelevant content before they can find the information that they require. This project attempts to solve this issue using a semantic web mining solution. The approach used in this project is one where we will scrape data from a number of websites and online forums, then structure it in an indexed manner and make it easily accessible through a website user interface. Patients would thus be able to easily access information pertinent to their health care issues and symptoms by simply entering keywords or phrases related to the disease and then viewing the results that the system presents.**

Keywords: SympGraph, MetaMap, Ontology, Apache Solr, Indexing, Scraping, HealthCare, Symptoms.

I. Problem Statement

Many health-related forums and websites provide valuable information on a wide range of medical conditions, treatments, and experiences. However, accessing and navigating these sources efficiently can be challenging due to the dispersed nature of the information and varying search functionalities across platforms. There is a need for a centralized platform that aggregates and organizes information from multiple health forums, allowing users to easily search for and access relevant content. The goal of this project is to develop a website that collects and indexes posts from various health forums and websites, focusing on discussions related to different medical conditions, symptoms, treatments, and medications. The key objectives of this project include:

- Scraping data from selected health forums and websites to gather posts related to various medical conditions and treatments.
- Building an ontology based on the relationships between diseases, symptoms, treatments, and medications mentioned in the scraped data.
- Indexing the collected data based on symptoms, diseases, and treatments to facilitate efficient search and retrieval.
- Developing a user-friendly interface that allows users to search for information based on their symptoms or medical conditions.
- Providing a map of diseases and their related symptoms to enhance understanding and navigation.
- Implementing a recommendation system that suggests treatments or alternative remedies based on user-generated content and forum discussions.
- Displaying posts with useful metrics to provide transparency in the recommendation system and help users make informed decisions.

By creating a centralized platform that aggregates and organizes health-related information from various sources, this project aims to provide users with a comprehensive and user-friendly resource for accessing relevant medical information.

II. Datasets

In our project, we aim to analyze the discourse surrounding various diseases, their symptoms and their treatments on discussion forums. These platforms offer a rich source of unstructured data in the form of user-generated content, including discussions about symptoms, advice-seeking posts, and community responses. To construct a comprehensive dataset for this analysis, we will employ web scraping techniques to collect data from a selected group of forums known for their active discussions on diseases. The forums selected for this study include:

- MedHelp: A user-centric health community where individuals share questions and experiences related to various diseases.
- Patient.info: An open forum for discussing a wide range of health conditions, capturing diverse user inquiries and responses.
- LiveScience Forums: A platform for scientifically inclined discussions on various diseases and their epidemiology.
- Centre for Addiction and Mental Health (CAMH): This forum focuses on mental health aspects related to different diseases.
- Mayo Clinic Connect: A patient community discussing recovery and management strategies for various illnesses.

For each forum, our data collection will target specific elements critical for our analysis, including post titles, content, author details, URLs, replies, sub-replies, and various forms of

user engagement metrics. The collected data will be stored in a structured format. This dataset will serve as the foundation for our project, enabling us to perform detailed semantic analysis and uncover insights into the public's perceptions, concerns, and knowledge sharing regarding diseases.

III. State-of-the-art methods and algorithms

A. SympGraph

SympGraph is a framework designed for extracting information from clinical notes using graphs that represent relationships between symptoms [11]. SympGraph examines the connections between symptoms in user posts and creates models based on these relationships. The process starts using the Metamap Tool, which extracts the medical and symptom concepts from user's posts.

Metamap is a Java-based tool developed by the National Library of Medicine. It is used for mapping the biomedical texts to UMLS Metathesaurus or identifying Metathesaurus topics within texts. It divides the extracted data into phrases and returns the detected concepts along with additional data.

B. Big Data Approach and Word2vec algorithm

The paper [4] introduces a method for searching and retrieving similar textual data within a large biomedical text dataset. This approach utilizes Word embedding models created with the word2vec algorithm, alongside a Big Data architecture for effective management.

C. Web Scraping Techniques And Tools

The paper [10] offers a comprehensive overview of web scraping methodologies, encompassing a range of techniques such as HTTP programming, HTML parsing, DOM parsing, computer-vision-based webpage analysis, and various web scraping tools like Mozenda, Visual Web Ripper, Scrapy, Web Content Extractor and Import.io.

IV. Research Plan

A. Phase 1: Gathering Data and Ensuring Ethical Practices

To kickstart our project, we will research various discussion forums and websites, expanding beyond Dr. Davulcu's initial list. These platforms host discussions on diseases, where users share queries, suggestions, and observations. Posts are interactive, allowing other users to respond and engage with each other. Once we select the right sources, we'll commence data scraping. Before diving in, we'll ensure our methods align with ethical standards such as GDPR [3] and HIPAA [2], implementing measures to protect user's privacy. After that, we'll extract key data elements like post titles, text, replies, and URLs related to the target disease. We'll utilise tools like Scrapy [1] and BeautifulSoup [9] for scraping, with Scrapy managing the web crawling and BeautifulSoup handling detailed data extraction. Additionally, we'll integrate Selenium [5] with Scrapy for scraping sites with dynamic content. We will also use modern techniques such as neural network-based web crawling for more efficient data gathering.

B. Phase 2: Data Processing, Cleaning, and Extraction

After gathering data, we'll focus on cleaning and preprocessing to eliminate noise and ensure high-quality data. Techniques like stemming [8], lemmatization [7], and stop word removal [6] will be used. Next, we'll use MetaMap framework to identify medical symptoms, diseases, and drug names, leveraging the Unified Medical Language System (UMLS) as a reference. We'll employ the SympGraph framework [11] to analyze symptom-disease relationships, identifying patterns to construct symptom-disease graphs. Finally, we'll train Word2Vec [12] models to capture semantic relationships between terms, enhancing our understanding of medical concepts.

C. Phase 3: Data Integration and Ontology Development

In this phase, we'll build an ontology to organize the relationships between diseases, symptoms, treatments, and medications. These ontologies will be structured hierarchically, with attributes such as disease name, symptoms, treatments, and medications. Using entity recognition and semantic analysis, keywords from text data will be mapped to these attributes. We will also incorporate modern approaches such as Graph Convolutional Networks (GCNs) to infer implicit relationships between medical entities from unstructured text data. We will simplify user searches through query parsing, semantic matching, and faceted search, streamlining information retrieval. Furthermore, we'll organize annotated data and ontologies in a relational database for efficient storage and retrieval, which will ensure seamless access to relevant information.

D. Phase 4: Search and Indexing

In getting the backend servers ready to handle requests from the upcoming frontend, we need to dig deep into SOLR to bring in and make sure we get the right data when users search. After this step, it's important to check if the search results are correct and useful. With Apache Solr, we'll organize comments and replies from discussion boards, use a co-occurrence graph to suggest symptoms and diseases to users, and create a system to find the best matching comments. By using Apache Solr, we can search based on the symptoms users give us, and show other relevant symptoms from popular posts.

E. Phase 5: User Interface Development

This phase of the project involves integrating all components into a user-friendly UI. This entails establishing a front-end to process user requests and display a prioritized list of posts, ensuring the most relevant response appears first. Moreover, the UI will incorporate a search feature to retrieve pertinent information from pre-defined structured data, leveraging the Apache Solr engine for efficient search and indexing functionality. Additionally, users will have the option to apply various filters to refine their search results further. If users pose queries, the system will provide accurate responses.

F. Phase 6: Evaluation and Refinement

In the final phase of our project, the system will be assessed using precision and recall metrics. Precision gauges the relevance of returned posts to a user query, while recall measures the inclusion of relevant posts in the system's output.

V. Evaluation Plan

The system will be evaluated using metrics for precision and recall. The Precision metric will be calculated based on the number of relevant posts that were returned by the system for a given query that was entered by the user. The Recall metric is calculated based on how many of the relevant posts related to the query were present in the system's output. Given below are the formulae for the two metrics:

Precision = # of relevant posts returned / Total # of posts returned for that query

Recall = # of relevant posts returned / Total # of possible relevant posts for that query

VI. Project Timeline: Tasks, Descriptions, Deadlines

TABLE I
Task Deadlines

Task	Description	Deadline
Researching and Scraping datasets	Identify relevant online discussion forums and websites hosting discussions on diseases, Utilize web scraping techniques to gather unstructured data from these platforms.	04/03/2024
Extract medicinal data	Extract medicinal data and annotate medicinal information such as symptoms, treatments, medications, and disease names using techniques like NLP, Metamap and SympGraph	08/03/2024
Ontology creation	Develop hierarchical ontologies to organize relationships between diseases, symptoms, treatments, and medications	15/03/2024
Indexing data with Apache Solr	Use Apache Solr. to implement indexing mechanisms and efficiently store and retrieve structured data	23/03/2024
Webpage/Interface creation	Design and develop a user-friendly webpage or interface for interacting with indexed data	05/04/2024

VII. Division of Work

TABLE II
Task Assignment

Task	Members
Researching and Scraping datasets	Manoj, Kshitiz, Pulkit
Extract medicinal data	Praveen and Aditya
Ontology creation	Gaurav and Pulkit
Indexing data with Apache Solr	Kshitiz, Gaurav, Aditya
Webpage/Interface creation	Manoj, Praveen

VIII. Acknowledgment

We want to thank Professor Hasan Davulcu for guiding us in figuring out what information and important parts we need for our project. His experience and knowledge helped us understand how our project can be useful and what it really needs to do. We are grateful for Professor Davulcu's expertise, which has been vital in shaping our project and ensuring its relevance to real-world needs. His support has been invaluable in our journey toward the success of this project.

References

[1] Scrapy: A Fast and Powerful Scraping and Web Crawling Framework. https://scrapy.org/.

[2] Health Insurance Portability and Accountability Act of 1996 (HIPAA), 1996.

[3] Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016.

[4] M Ciampi, E Masciari, G de Pietro, and S Silvestri. A big data approach for health data information retrieval. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2533–2540, San Diego, CA, USA, 2019.

[5] Selenium Contributors. Selenium: Web Browser Automation. https://www.selenium.dev/.

[6] Julie B. Lovins. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, 11(1-2):22–31, 1968.

[7] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[8] Martin Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, 1980.

[9] Leonard Richardson and Jonathan X. Z. Shaw. Beautiful Soup: Python Library for Web Scraping. https://www.crummy.com/software/BeautifulSoup/.

[10] Anand V Saurkar, Kedar G Pathare, and Shweta A Gode. An overview on web scraping techniques and tools. 2018.

[11] Parikshit Sondhi, Jimeng Sun, Hanghang Tong, and Chengxiang Zhai. Sympgraph: A framework for mining clinical notes through symptom relation graphs. In *KDD'12 - 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1167–1175, September 2012. 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2012 ; Conference date: 12-08-2012 Through 16-08-2012.

[12] Illhoi Yoo, Jinbo Bi, and Xiaohua Hu, editors. *2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019, San Diego, CA, USA, November 18-21, 2019*. IEEE, 2019.