Mining Healthcare Websites

Hari Sai Charan Challa
Department: Computer science
Arizona State University
Tempe, United States
hchalla2@asu.edu

Abdulla Mohammed Nihad
Department: Computer science
Arizona State University
Tempe, United States
anihad@asu.edu

Aditya Mettu
Department: Computer science
Arizona State University
Tempe, United States
amettu3@asu.edu

Sandeep Kallepalli
Department: Computer science
Arizona State University
Tempe, United States
skallep2@asu.edu

Vikas Kamineni
Department: Computer science
Arizona State University
Tempe, United States
vkaminen@asu.edu

***Abstract*—Healthcare forums provide a wealth of information on diseases, treatments, and medications, frequently conveyed through personal stories. Yet, this important information is usually available in an unstructured, natural language form. The absence of organized data, especially concerning the connections among diseases, symptoms, and medications, presents a major obstacle in extracting valuable insights.**

We are dedicated to solving this problem by gathering and scrutinizing data from healthcare sites, particularly looking at symptoms and associated medications. We intend to compile a list of potential diseases based on specified symptoms, and conversely, list possible symptoms and treatments for identified diseases. Our objective is to analyze the wealth of medical data shared by individuals to spot trends and patterns within the healthcare sector. This effort is expected to support more knowledgeable decision-making in the pharmaceutical industry, thereby enhancing the decision quality related to healthcare. Moreover, our research includes examining the Covid virus (SARS-CoV-2) to address the concerns of those affected by the pandemic or exhibiting similar symptoms.

***Keywords*–Healthcare mining, Covid, Web Scraping, SympGraph, Metamap, Apache Solr, Indexing, Ontology, UMLS, Page Ranking**

I. PROBLEM DEFINITION

In the dynamic field of healthcare, gathering and combining information from different sources is essential for progress in medical studies, identifying diseases, and developing treatment plans. Our initiative tackles the important task of collecting a wide range of data from healthcare websites and forums. These online spaces are filled with an abundance of unstructured details on illnesses, symptoms, and treatment options. The aim is to distill valuable understanding, trends, and wisdom from this unstructured information, establishing a foundation for smart health-related decisions. Our project is structured around the following process:

- Utilizing sophisticated web scraping methods to collect unstructured information from healthcare websites, which includes various viewpoints on diseases, symptoms, and treatments [2].
- Analyzing the gathered unstructured data with MetaMap, which works by linking free-text medical data to concepts within the Unified Medical Language System (UMLS) [3]. MetaMap conducts semantic annotation, associating concepts and medical terms with their respective UMLS identifiers, thus transforming varied and unstructured medical terminology into a uniform and organized format. This organized data is then saved in a database for effective retrieval of information.
- Using an advanced page ranking algorithm to find and present the most relevant web pages linked to the disease entered by the user [6]. In a similar manner, it shows symptoms related to a specific disease and vice versa.
- Creating a user interface that allows users to enter a disease name or symptom and easily view related results. Enhancing the user experience through the integration of intelligent symptom suggestions via SympGraph, streamlining the search process [1].

II. DATA SETS

Numerous online data sources exist for scraping pertinent information. We have utilized a variety of sources and forums in our research to gather data and present the content. We intend to use the following websites:

a) Medhelp: Includes a collection of questions, their posting dates, responses, and ratings of helpfulness.
b) Patient.info: Offers an extensive array of health-related details and resources, covering a wide spectrum of symptoms and diseases.
c) Livescience.com: Features a list of questions, response, comment count, views, reaction score, likes and points.
d) Camhx.ca: Specifically a COVID-19 discussion platform, it provides a compilation of questions, replies, dates, and moderator responses related to COVID-19.
e) Mayoclinic.org: Consists of diseases, their descriptions and list of topics, responses, reactions and date.

The data, once scraped, includes attributes such as subject, URL, content, author, responses, and sub-comments. This unstructured data is then transformed into a structured format for subsequent analysis.

III. STATE-OF-ART METHODS ALGORITHMS

We have discovered several advanced techniques for gleaning important insights from healthcare websites, with SympGraph being one of these innovative approaches [1]. SympGraph is a complex data mining framework developed specifically for mapping and analyzing the links between symptoms identified in clinical notes. It seeks to address the issues arising from unstructured clinical notes in electronic health records (EHRs). The fundamental concept of SympGraph is the formation of a graph structure, where symptoms are illustrated as nodes, and their connections or co-occurrences are depicted as edges. This approach is aimed at understanding the intricate relationships among symptoms recorded in clinical documentation. This graphical model is created automatically by extracting symptom data from a vast array of patient clinical notes. The visual representation that emerges provides crucial insights into the relationships between various symptoms, facilitating the discovery of patterns and linkages crucial for activities such as disease diagnosis, treatment formulation, and other medical processes. A key feature of SympGraph is the 'symptom expansion' function. This feature leverages the structure of the graph to broaden an initial symptom set to encompass additional, related symptoms. Such an expansion is immensely beneficial in healthcare, enabling the identification of further symptoms that may be significant for a patient's diagnosis or assisting in the investigation of possible links among different symptoms. To conclude, SympGraph is a dynamic and effective instrument for extracting clinical information from electronic health records, especially useful for navigating through unstructured clinical notes. With its novel methodology in mapping and scrutinizing the connections between symptoms, SympGraph provides researchers and medical practitioners the tools necessary to achieve a more profound comprehension of disease trends, uncover novel treatment paths, and ultimately improve patient care and health outcomes.

The National Library of Medicine has developed a specialized instrument known as MetaMap UMLS (Unified Medical Language System) [3]. This tool is designed primarily for processing natural language in clinical texts, with a focus on identifying and categorizing clinical concepts from unstructured sources such as clinical notes, medical articles, and various health documents. MetaMap UMLS leverages an extensive database of medical terms and concepts, the UMLS Metathesaurus, which contains more than 3 million biomedical concepts and over 11 million unique concept names. Additionally, it employs a combination of heuristics and algorithms to recognize medical terms and to link related concepts effectively. The result of processing by MetaMap UMLS is a series of connections established between medical terminology in the input text and relevant concepts in the UMLS Metathesaurus. These connections are pivotal for various clinical applications, such as identifying groups of patients with specific conditions, tracking disease progression, and forecasting clinical outcomes. Overall, MetaMap UMLS emerges as a powerful tool for analyzing clinical texts. It enables researchers and healthcare providers to extract meaningful insights from unstructured data, which might otherwise be difficult to interpret. Leveraging the vast repository of the UMLS Metathesaurus, MetaMap UMLS excels in pinpointing and elucidating medical concepts within texts, thus improving the accuracy and depth of clinical data analysis [3]. Apache Solr is an open-source search platform built for enterprise-level applications [4]. It provides numerous advanced search features, such as full-text search, hit highlighting, and faceted search. Developed on the Apache Lucene search engine library's architecture, Solr excels in speed and scalability, accommodating a wide range of uses. Its versatility is one of Solr's key attributes, enabling it to index and search through various document types, including XML, JSON, and binary formats like Microsoft Word and PDF. Additionally, Solr offers sophisticated search capabilities, such as faceted search, which allows users to narrow down search results through categories or facets, and hit highlighting, which draws attention to search terms found within the results. Solr has established a strong presence across multiple sectors, such as e-commerce, healthcare, and finance, where effective search capabilities are essential for analyzing data and making informed decisions. It is highly regarded by organizations for its scalable, fast, and flexible nature, which positions it as the go-to option for integrating advanced search features into applications. In essence, Apache Solr acts as a powerful and versatile search platform, enabling businesses and developers alike to embed rapid and scalable search functionalities into their applications [4].

IV. RESEARCH PLAN

The goal of this project is to create an improved information retrieval system focusing on disease data obtained from healthcare websites. This system will enable users to effortlessly search for and access valuable information about different diseases, such as symptoms, treatments, and user experiences, via an intuitive interface. To accomplish this objective, the project will be structured into four separate phases, with each phase expanding on the work of the preceding one.

A. Data Collection

The initial stage of this project involves an extensive review of healthcare forums and websites to pinpoint suitable sources rich in discussions about various diseases. Sites like MedHelp, Patient.info, WebMD, and Drugs.com, among others, will be targeted for data extraction. Our efforts will concentrate on capturing discussions encompassing a broad range of diseases to compile pertinent data. For the collection of this data, we will utilize web scraping tools like Scrapy [5].

B. Data Structuring and Ontology Creation

After gathering the data, it will be organized and stored in JSON/CSV format. To tackle the issue of synonyms for diseases, symptoms, and treatments, we plan to employ the UMLS Terminology and MetaMap for extracting medical meanings from the JSON/CSV data [3]. We will develop ontologies to define the connections among symptoms, diseases, and treatments. These ontologies will then be saved in a relational database, with MetaMap's results being incorporated into this database. The use of these ontologies will enable the delivery of pertinent information through the front-end user interface.

C. Data Indexing and Recommendation

In the third phase, the data we've collected, encompassing comments and replies, will be indexed using Apache Solr [4]. We will conduct further investigations to enhance the efficiency of the indexing process. A ranking algorithm is to be devised to retrieve the top N matching comments in response to user queries [6]. The prominence of symptoms and medications mentioned in various posts will be analyzed to determine the occurrence rate of certain cases. By evaluating the significance of associated symptoms and the frequency of keywords within posts, we aim to display content that is most relevant to users.

D. User Interface Development

In the final phase, we'll build a graphical user interface (UI) for easy and clear information access. Users can search data with specific parameters to find relevant information and customize their search results to explore topics on various diseases.

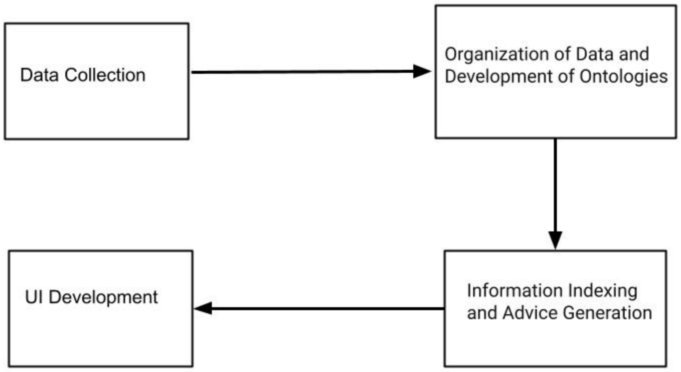
Data Collection
Organization of Data and Development of Ontologies
Information Indexing and Advice Generation
UI Development

Fig. 1: Project Plan Flowchart

The primary aim of this project is to create a useful tool for those looking for precise and pertinent disease information from healthcare websites. This system will provide an intuitive interface, allowing users to efficiently search and explore data, thereby enhancing the understanding of different diseases.

V. EVALUATION PLAN

The evaluation strategy for the system will center on its efficiency in presenting users with pertinent posts. To measure its success, we'll employ Precision and Recall metrics. Precision, indicating the ratio of relevant posts within the returned results, will be determined by dividing the number of relevant posts returned by the total posts returned. Recall, assessing the system's capacity to fetch all relevant posts, will be determined by dividing the number of relevant posts returned by the total number of relevant posts available. Analyzing both metrics will offer a detailed insight into the system's effectiveness in not just retrieving posts but ensuring their relevance to users' queries and interests. This evaluation approach will shed light on the system's performance, informing enhancements to improve user satisfaction.

VI. PROJECT TIMELINE AND TASK DIVISION

Tasks	**Assignee**	**Deadline**
Gathering and extracting information from the specified source websites	Hari	03/06/2024
Employing the MetaMap text annotation tool to identify medical concepts within user-generated content and reviews	Vikas	03/13/2024
Developing structured databases and backend infrastructure	Aditya	03/13/2024
Implementing Data Indexing and Recommendations with Apache Solr	Sandeep	03/20/2024
Developing Frontend User Interfaces	Nihad	03/27/2024
System Integration and Testing	Everyone	04/05/2024

Table I : Project Timeline and Task Division

REFERENCES

[1] Parikshit Sondhi, Jimeng Sun, Hanghang Tong, ChengXiang Zhai, *SymGraph: A System for Analyzing Clinical Notes Using Symptom Relationship Graphs*. KDD 2012: 1167-1175.

[2] Anand V. Sawarkar and Kedar G. Pathare and Shweta A. Gode, *A Guide to Techniques and Tools for Web Scraping*. (2018).

[3] A. Aronson, *MetaMap: A Utility for Identifying UMLS Concepts within Textual Data*. (2018).

[4] Apache Solr: A robust, open-source search platform that provides distributed indexing, search, and analytics capabilities.

[5] Scrapy: A collaborative, open-source framework designed for efficient data extraction from websites. It's built to be quick, straightforward, and highly extendable.

[6] Exploring Graph Data Science: An Introduction to the PageRank Algorithm.