HEALTH CARE DATA MINING

Dhiren Tejwani, Kaushik Sampath, Rahul Arora, Raj Buddhadev, Riddhi Patel, Rishab Banerjee, Tithi Patel

OVERVIEW

The proposed project aims to show relevant symptoms for a given disease and find the conversation threads for a disease or symptom and sorts them by maximum relevance. Sites like WebMD and diabetes.co.uk will provide the dataset (conversation threads) that is needed for the implementation of the project. Since there are several available forums which offer discussion on many diseases, the project tries to provide a filter-based interface, where a user can conveniently look for discussions. The major challenges in the project include identification of health pages that have enough information, extraction of symptom related information on the web pages, identification of common symptoms, ensuring search is intelligent, and storage and presentation of information.

DATASETS

We will mainly be using **WebMD**, a trustworthy source for health and medical news and information for our dataset. We will be scraping the discussions from the forum of WebMD and use those for our project.

We will also be scraping the data from **diabetes.co.uk**, which is operated by Diabetes Digital Media and serves as a platform for diabetic patients to learn more about the disease and discuss their concerns with different users on the forum.

There are a few databases like **healthdata.gov** and **Medline Plus** also available to the public where we can find information on certain diseases, relevant symptoms and their treatments. We might also use these databases to get more information for our project.

EXISTING STATE-OF-ART

The prevalent state-of-the-art methods in the field of medical convoyed with machine learning has a wide range. Major emphasis has been laid on the Image Segmentation. The research areas

pertaining to image segmentation and the ones which have seen a major development are Vessel Segmentation, Cell Detection, and Nuclear Segmentation.

In the last two decades, a variety of different ML techniques and feature selection algorithms have been widely applied to disease prognosis and prediction. The vast majority of publications makes use of one or more ML algorithms and integrates data from heterogeneous sources for the detection of tumors as well as for the prediction/prognosis of a cancer type. The state-of-the-art areas include Breast Cancer Detection, Lung Cancer Diagnosis, Skin Cancer Classification and the like.

While, there have been many interesting works in the domain of healthcare that uses the power of data to provide a useful analysis of symptoms, diseases, and its detection. However, there isn't much recognized work done when it comes to discussion forums in the healthcare domain. Yes there are few discussion forums for specific as well as general diseases, but there aren't any user-friendly filtration operations applied on those discussions to extract out the more relevant discussions from the forum. Hence, through our project we will try to build an interface that will help the user to find out the discussions (scraped from different forums) related to their specific symptoms and diseases.

CHALLENGES & RESEARCH PLAN

1. Identifying Information-rich healthcare web pages

This step addresses the web crawling and web searching problems of semantic web mining. A web crawler is an application that scans the world wide web in an automated manner to categorize and pull information on the basis of user needs. A customized crawler can be used to specifically search and index healthcare information sites providing information about specific diseases. Next we need a web page ranking and searching mechanism to identify the top sites containing information about diseases.

The critical challenges identified in the above process is studying and identifying the best web crawling techniques among the many available techniques to identify the best web pages containing information about symptoms for a given disease. Alternatively, we may use an efficient search engine API to identify the best web pages presenting information about disease, in addition to the ones already identified manually in the datasets section. Once the content rich web pages are identified, the major difficulty would be converting the unstructured HTML data to a semi structured format like JSON or XML and then using a customized parser to extract textual records describing the symptoms

2. Extracting Symptom related information from the web pages

The next semantic web mining challenge to be addressed is information extraction from web pages. A customized web page parser is required which can identify certain keywords and can extract textual information from the unstructured HTML web pages about the symptoms related to the disease in question. There are novel approaches like Ontology Based Information Extraction. Different information extraction techniques must be analyzed, and the best method needs to be identified and used to extract the symptoms.

Software's like Semantic Tagging may be used to better label html documents extracted from the documents. It helps in facilitating communication and finding information. When we add semantic tags to blog posts or documents, we are actually providing more information about the post. For example, <h1> tag which is usually the title of most blog posts indicates that the enclosed text is a headline 1. This is semantic as well as presentational because both the user and the browser know that it is a headline tag.

Critical Challenges:
How will we expand newly mined symptoms, without duplicates? For instance, High Fever and Hyperthermia are synonyms for the same symptoms and not two different symptoms for the same disease.
Should specific attributes about a disease symptom be extracted which are important and required to uniquely describe that symptom? While 'heartburn that occurs suddenly' can be a symptom for 'Heart Attack' (myocardial infarction), 'regularly occurring heartburn after meals' can be a symptom for 'Peptic Ulcer'.
How do we identify co-occurring symptoms for a disease? How do we represent specific relations between symptoms which are important to differentiate diseases?

3. Making searches intelligent

Boosting techniques for example Elasticsearch boost can be used to extract important terms from the search string to make searching more effective. For example, if someone searches "leukemia cancer", the term "leukemia" should be boosted as "cancer" will be present in multiple documents, but documents with "leukemia" should be given higher priority-- "leukemia" will be deeper down in the ontology tree, thus, implying a more specific search term.

4. Storing and Presenting the Identified Symptoms

The most effective and efficient way to store the diseases and symptoms must be discovered. For our application, we may either use a SQL relational schema or a No-SQL DB. Best fit of the chosen storage platform for our application and requirements must be

analyzed. Also, for the proposed system, an intuitive UI must be designed so that a user can input a disease and our application must be capable of mining the relevant symptoms from the web and displaying the results in an organized manner to the user. The search can be made for effective by storing the data and indexing using Elasticsearch.

IMPLEMENTATION TIMELINE

Time Period	Tasks
20th Feb - 5th Mar	Data Scraping, Data cleaning & processing, Developing user interface
5th Mar - 20th Mar	Applying algorithms on data, Putting data on the interface
20th Mar - 5th Apr	Applying filtration logic and other user-friendly operations
5th Apr - 20th Apr	Final report, code and other supplementary materials

WORK ALLOTMENT

Name	Work Allocated (Tentative)
Dhiren Tejwani	Data Scraping, Applying algorithms on data, Applying filtration logic
Kaushik Sampath	Data Scraping, Applying algorithms on data, Applying filtration logic
Rahul Arora	Data cleaning & processing, Applying algorithms on data, Applying filtration logic
Raj Buddhadev	Data Scraping, Applying algorithms on data, Applying filtration logic
Riddhi Patel	Develop User Interface, Putting data on the interface, Integrating algorithms within the interface
Rishab Banerjee	Data cleaning & processing, Applying algorithms on data, Applying filtration logic
Tithi Patel	Develop User Interface, Putting data on the interface, Integrating algorithms within the interface