International Institute of Information Technology, Hyderabad

Project Title
[System for Diagnosing Health]

Team Members

Name	Roll number
Sushant Kumar Thakur	201101095
Sindhu Kiranmai Ernala	201156078
Poorva Bhawsar	201305555
Chaitanya Hemant Kiran Kumar Boda	201205602

Table of Contents

Problem Statement

Given some initial symptoms detect the possible disease efficiently.

Project Description

In this project we aim to build a system providing convenient access to knowledge about behavioral factors involved in human diseases, as well as body parts and symptoms that are affected and caused by these diseases. The system should be capable of automatically extracting relations between the symptoms given by user and the probable diseases.

Motivation

Today most of the medical information, is not readily available to the public. Approaching the doctor is the only means of being informed about the illness. However, if this knowledge could be spread to the population, many diseases could be prevented, diagnosed earlier and more accurately, and thus treated better, and cured more effectively; even epidemics and pandemics could be avoided. There are also many generic symptoms and casual doubts which are not readily taken to the doctor for consultancy. In such cases, our system helps the user by providing relevant information. Thus, our main motivation lies in providing accessibility of the medical expertise for the populace.

Approach

- **Crawling:** Initially, we shall crawl WebMD website to obtain structured/ unstructured data in html format.

- **Parsing and Extraction**: Next step is to parse the dataset to find the patterns that relate symptoms and diseases using entity extraction and relation extraction that are based on text mining and pattern based extraction. We represent Symptom and Disease as the Entities and name the relation between them as 'isSymptomOf'.
  For example from the text “Coughing is a symptom of Asthma” , we extract Asthma(disease), Coughing (symptom) as entities and the entity-relation will be shown as 'Coughing isSymptomOf Asthma'.

- **Indexing**: We build the inverted index with symptoms as the dictionary terms and set of probable diseases as the posting list. Each disease in the posting list is ranked by the probability of symptom's prominence in the disease.

- **Retrieval**: Once user provides the set of symptoms, corresponding posting lists are fetched , AND / OR operations are performed on the posting lists and the resultant set of diseases is displayed to the user in ranked order.

Similar approach can be followed to include the causes and remedies for the diagnosed disease.

Dataset

We had two choices for our dataset,

1. WebMD (http://www.webmd.com/default.htm)
2. Mayoclinic (http://www.mayoclinic.org/)

   We chose to proceed with the WebMD dataset as it has a richer knowledge base and well defined, structured data which makes it easy for information extraction. This website also provides various combinations of symptoms with relevant diseases.

Final Deliverable

We aim to build a system which primarily gives most probable diseases for a given set of symptoms. The system should be user friendly with interfaces where in the user can select the symptoms from a predefined list, categorized with respect to body parts. Further we aim to extend it to provide causes and also remedies for the diagnosed disease.

References

Research Paper: Text Mining for Building a Biomedical Knowledge Base on Diseases, Risk Factors, and Symptoms
(https://domino.mpi-inf.mpg.de/intranet/ag5/ag5publ.nsf/0/F8C50D2AB9FB5FEFC12579420049025F/$file/MinYe_Master%20Thesis.pdf)

Website:
https://uts.nlm.nih.gov/home.html

End of Scope Statement