Search Mechanism focused on Diseases, Symptoms and Treatments

André Esteves, Francisco Filipe, Helena Montenegro, Juliana Marques
Information Description, Storage and Retrieval DAPI
Master in Informatics and Computing Engineering MIEIC
Faculty of Engineering, University of Porto FEUP
Porto, Portugal
{up201606673, up201604601, up201604184, up201605568}@fe.up.pt

***Abstract*—Nowadays, there are several search engines publicly available for people to use in a global scale, containing information about almost anything we can think about. However, when it comes to health matters, these search engines might not be the best choice. Very frequently there will be documents from unreliable sources containing misleading information that might wrongly worry the average user.**

From the data sources selection to the data retrieval, cleaning and enrichment, this article focuses on detailing all the main stages of building an health related search mechanism that tackles the problem of unreliable information very commonly found in today's search engines. A system where all the information is trustworthy and the user needs not to worry about its contents not being accurate.

***Index Terms*—Diseases, Symptoms, Treatments, Drugs, Specialities, Causes, WikiData, Wikipedia, Search Systems**

I. Introduction

Search engines such as Google are used by billions of people on a daily basis, to retrieve information about various subjects, including health matters. When a person is worried about a symptom and searches for it on the web, many documents are retrieved, including some from unreliable sources, such as Yahoo answers where someone asks a question and any person can answer, for example. The average person does not have the knowledge to discern between reliable and unreliable sources of information, easily believing any information that comes up even when the sources are not specified, just because the content seems legitimate. When it comes to health, unreliable and exaggerated information can lead to anxiety and panic in patients.

There are search engines that specialize in obtaining health information, such as MedWorm [5], however, these engines focus on retrieving biomedical texts and articles published in the health community, which are useful for health specialists, but not for the average user, due to difficulties in understanding the content of the documents and unfamiliarity with the technical terms used in these. A normal user needs a simple and straightforward mechanism that, rather than focusing on advances in the medical community or overly specific documents, shows information about diseases and respective symptoms and treatments.

As there is currently no search mechanism that allows a person to easily obtain reliable information about health matters, the goal of this project is to develop one, focusing on diseases, treatments and symptoms.

This article serves to describe the first step in this project, which regards the preparation of the data set. Section II describes the domain and conceptual model of the problem. Section III explains the pipeline implemented for the process of preparing and characterizing the data. The final section is focused on the system results and retrieval tasks.

II. Conceptual Model

The conceptual model (Fig. 1) represents the entities of interest and the relations between them. Note that this model applies to the data and also to the domain since all entities of interest and their relations are depicted.

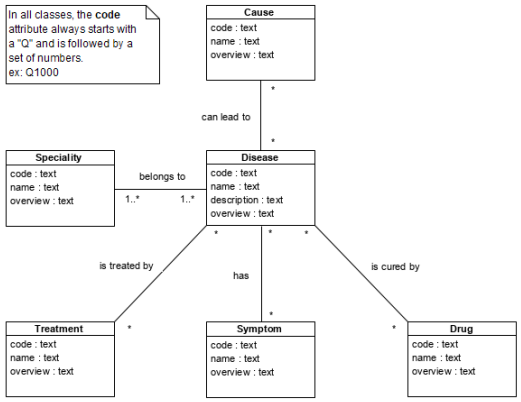

Fig. 1. Conceptual Model

The Disease class is the core of our data set and is related to all the other classes of the domain. This class stores all the relevant information about a disease and it's attributes are:

- **code** - a unique attribute that identifies a disease.
- **name** - the disease's name.
- **description** - a small text that describes the disease.
- **overview** - a detailed text about the disease extracted from Wikipedia.

Furthermore, our domain has 5 more classes, all related to the Disease class.

- **Cause** - Class that represents all the causes that can lead to disease.
- **Speciality** - Class that represents all the specialities associated with a disease. All the diseases have at least one speciality.
- **Treatment**- Class that represents all the treatments that can be used to treat a disease.
- **Symptom** - Class that represents all the symptoms a disease can have.
- **Drug**- Class that represents all the drugs that can be used to cure a disease.

All these classes have 3 attributes, which are:

- **code** - a unique attribute that identifies the instance.
- **name** - the instance's name.
- **overview** - a detailed text about the instance extracted from Wikipedia.

III. Data Preparation

For the process of preparing the data, we developed a pipeline (Fig. 2) divided in 5 processes: data collection, storage, cleaning, enrichment and characterization.

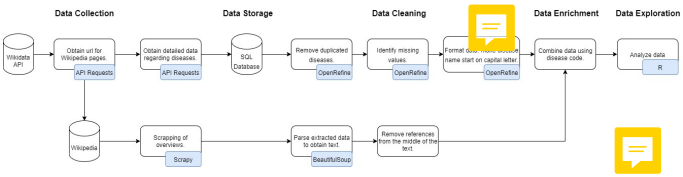

Fig. 2. Data pipeline diagram

A. Data Collection

The data set was obtained from two sources: Wikidata and Wikipedia.

*1) **Wikidata**:* Wikidata is a large database with structured data, containing information of various subjects, including diseases. This data can be copied, modified, and distributed, even for commercial purposes, without needing permission, under the license "Creative Commons Public Domain Dedication 1.0" [1]. However, information in Wikidata can be modified by anyone, without needing to be verifiable against authoritative sources and is, therefore, unreliable. Furthermore, after analyzing the obtained data we arrived at the conclusion that it is severely lacking, with many diseases that do not have information about symptoms or even health specialties.

*2) **Wikipedia**:* Wikipedia is an encyclopedia that contains unstructured textual data which is free to be shared and adapted for any purpose, including commercial purposes, as long as appropriate credit is given, under the license "Creative Commons Attribution-ShareAlike 3.0 Unported" [2,3]. Although anyone can edit Wikipedia pages, Wikipedia has a policy that states that any alterations or additions must be verifiable against an authoritative source, which makes it reliable as a data source [4].

Data from Wikidata was obtained in json format through its API. Along with information about the diseases and respective characteristics, we also obtained the url for the respective Wikipedia page, which allowed us to obtain overviews of diseases, symptoms, treatments and of the other classes, through crawling and scraping, using the tool Scrapy.

The use of Wikidata, which possesses unreliable and incomplete data, is a limitation that risks the reliability of the search mechanism being developed.

B. Data Storage

We stored the data collected in an SQL database, using Microsoft's Azure SQL Database [6].

C. Data Cleaning

The data extracted from Wikidata contained a lot of diseases without any connections to other classes, such as symptoms, treatments, and so on. We made the decision to remove all diseases that contain less than two connections to other classes. After that, we removed all symptoms, treatments, health specialties, causes and drugs that were not connected to any remaining disease. This was achieved by performing DELETE statements on the database in Azure Data Studio. We also capitalized the first letter of the name of diseases, symptoms, treatments, causes, health specialties and drugs, using UPDATE statements on the database. By making the code of the diseases retrieved from Wikidata unique, we ensured that there were no duplicate values in the database.

The scraped data obtained from Wikipedia was in html format. The data was parsed using the tool BeautifulSoup to extract the text. We then proceeded to remove special characters and references, using python.

D. Data Enrichment

The data from Wikidata was enriched with data from Wikipedia. The data was easily joined, using the code of diseases and of the other classes, with UPDATE statements on the SQL database.

E. Data Characterization

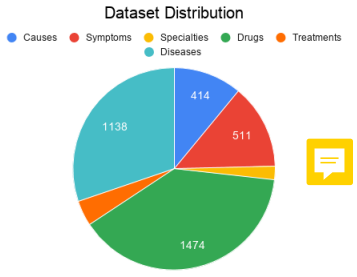

Fig. 3. Pie chart with data set distribution

The data set distribution (Fig. 3) shows that the largest classes present are **Drugs** (approximately 40%) and **Diseases** (approximately 30%). The smallest class is the one that concerns **Specialities** (approximately 0,02%).

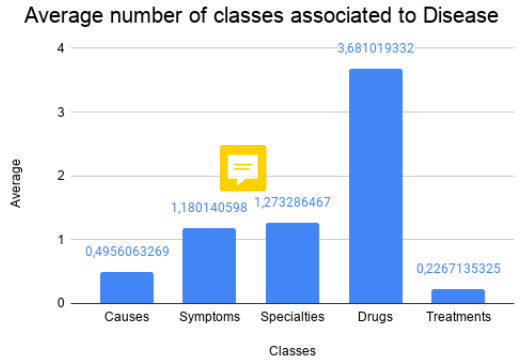

Fig. 4. Bar graph with average number of classes associated to disease

The average number of classes associated with a Disease graph (Fig. 4) shows that there are less causes and treatments associated to diseases than the remaining classes. There is more information about drugs on diseases, since a disease has, on average, around 3.7 drugs.

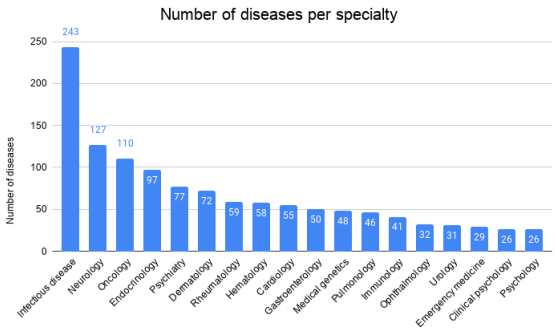

Fig. 5. Bar graph with number of diseases per specialty

When it comes to the number of diseases organized by specialty (Fig. 5), it's possible to see that the specialty associated with the most diseases is **infectious diseases** (243 diseases), and the specialty with the least number of diseases is **psychology** (26 diseases).

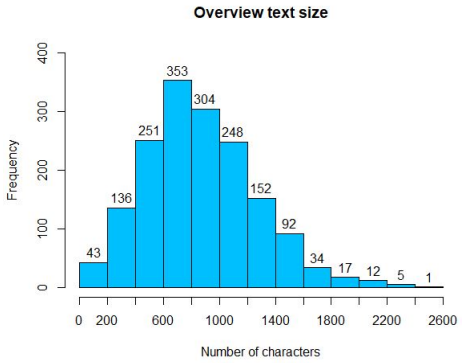

Fig. 6. Histogram with overview text size

When it comes to the size of the overview texts (Fig. 6) of the documents, most of the overviews have about 600-1200 characters (0,56 %). About 10% of the overviews have more than 1400 characters and only 0,3% have less than 200 characters.

IV. System

A. System Results

There will be 1648 documents in the system, distributed in three classes: Disease, Treatment and Symptom, as shown in Fig. 7.

1) ***Disease****:* The document Disease will contain a small overview of two paragraphs about the disease and a list of symptoms, treatments, causes, drugs and health specialties associated with the disease, if these exist. The user will be able to search for the disease by keywords associated with it, present in the overview, and by symptoms, treatments, causes and health specialties.

2) ***Treatment****:* The document Treatment contains an overview of the treatment and a list of diseases that can be cured by the treatment. A user will be able to search for the treatment by the name of a disease or by a keyword present in the name or overview of the treatment. Not all treatments in the system will be a document, only the ones that contain overviews.

3) ***Symptom****:* The document Symptom contains an overview of the symptom and a list of diseases that can have the symptom. This document, similarly to the treatments, should appear when a user searches by a disease or a keyword present in the name or overview of the symptom. Not all symptoms in the system will be a document, only the ones that contain overviews.

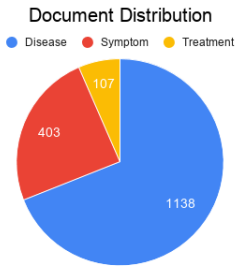

Fig. 7. Pie chart with document distribution

B. Retrieval Tasks

The expected return value of the retrieval tasks are documents which represent diseases, treatments and symptoms as these are the focus of the project. These documents will be retrieved based on the title's content and in the respective overview's content.

Other possible retrieval tasks include:

- Retrieve disease by symptom or health specialty.
- Retrieve treatment by disease.
- Retrieve symptom by disease.

V. Conclusion

The purpose of this project was to create a search mechanism that allows a person to easily obtain reliable information about health matters, focusing on diseases, treatments and symptoms.

In this first stage we successfully populated an Azure SQL database by combining information retrieved from both the Wikidata and Wikipedia. This information was later cleaned and enriched resulting in a total of 1648 documents distributed amongst Diseases, Treatments and Symptoms.

Having now a largely populated database, the next step is to make use of its contents resorting to information retrieval tools and free-text queries.

References

[1] Wikidata: Introduction, https://www.wikidata.org/wiki/Wikidata:Introduction

[2] Wikipedia: Text of Creative Commons Attribution ShareAlike 3.0 Unported License, https://en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License

[3] Creative Commons - Attribution ShareAlike 3.0 Unported License, https://creativecommons.org/licenses/by-sa/3.0/

[4] Wikipedia: Verifiability, https://en.wikipedia.org/wiki/Wikipedia:Verifiability

[5] Medworm: Medical Search Engine and RSS News, https://medworm.com/

[6] Azure SQL Database, https://azure.microsoft.com/en-us/services/sql-database/