CS 410 Tech Review
Akarsh K Bhagavath
Akarshb2
06/11/2022

Introduction

In the class we learnt various processes and methods for text retrieval and text mining. For my tech review, I wanted to go through the current state of the processes/ tools available to people interested in text retrieval related to healthcare applications. I believe text retrieval in healthcare/ text analytics on healthcare is a budding industry because of the following reasons. In the United States, I learnt from my Deep Learning healthcare class that healthcare is a $2.3 Trillion dollar industry, however, only $900 billion of that is used efficiently. High amounts of man hours and system resources are used up because patient data is unstructured across various medical providers, and finding a linear patient history, with clean data and less clutter is hard to come by. This is because a singular patient can have multiple healthcare providers throughout the span of their life, and the records passed to those medical providers, if at all, can be highly unstructured. As such, there is a need for large engineering effort in the healthcare industry to streamline the process of retrieving medical information for a patient, and this information needs to be consistent, accurate, and helpful to future medical providers assisting the patient. With a more streamlined and unified effort to mine/retrieve text data, there will be lesser wastage of money, resources, and man hours creating repetitive health records for the same patient, and those being lost/ mishandled by healthcare providers.
For this reason, I chose to read about Information Retrieval for Healthcare by William R Harsh of Oregon Health and Science University. This tech survey will contain a short description about the contents of the chapter, and personal opinions/ comments about it.

About the Chapter (Body)

The author starts by talking about the basic information retrieval process universal to all types of information retrieval. We understand that before retrieval, data needs to be curated and structured. Once structured, the data can then be extracted into a metadata format to reduce multiple words into one. Finally, this metadata can be used by search engines, queries or placed into other items (14.1). It is interesting to see here that the author has left out the process of curating the healthcare data in the first place. Data curation is an important step in creating healthcare data needed for information retrieval. In my opinion, There is a plethora of healthcare data lying around in various parts of the web that has not been curated properly for information retrieval. There is also a lot of medical data not yet converted into online copies, and hence cannot be retrieved using conventional text retrieval techniques. I believe the author should have included a section in the beginning talking about the need for healthcare data curation before talking about the information retrieval aspect.
We are then introduced to the various kinds of information present in healthcare data, and how healthcare providers seek it (14.2.1). The author mentions the various kinds of information present in healthcare data such as diagnosis/ procedures/ insurance information etc, and how this information can be retrieved by healthcare providers (healthcare provider can knowingly query for it or healthcare provider can unknowingly query for it through a related search).

The author then goes on to mention the various information present in these medical texts (14.3). We come to learn that there are some information that is critical to be matched exactly, like annotated content, and bibliographic content, while the full text body of the medical text cannot be matched exactly. Rather we look for keywords in the text body and match/rank it with the query appropriately. This is done using Term Frequency, and Inverse Document Frequency to prevent stop words skewing the search. We can see here a clear association between healthcare data and text retrieval methods I learnt in class, which peaks interest.

We then are introduced to the topic of indexing in healthcare (14.4). From previous classes, I am aware that indexing is the process of associating metadata with topics. We see that certain kinds of metadata can be automatically assigned to contents/topics (e.g. a simple indexer can associate a heart attack to heart disease), but sometimes, human intervention is required to manually index the metadata to topics (e.g. a simple indexer may not know to associate heart attack with myocardial infarction, or the KRAS biomarker with lung cancer. This needs more domain knowledge about the metadata and the content). I personally thought that the author could have gone into more detail about meshing the two types of indexing together. While we have automated indexing for simple types of indexing and manual indexing for complex topics, we could mesh the two together and create a filtering indexing system, such that the automated indexer takes a first pass at the metadata and assigns a topic to it, along with a confidence score. If the confidence score is low, then the metadata can be reassigned to a manual (human) indexer with domain knowledge to index the metadata into a more niche topic.

The author then talks about the core of the retrieval process (14.5). As mentioned before, we see that certain parts of the medical text need to properly matched, while other parts of the text only need a partial matching with a score to keep track of how relevant it is. In accordance we see Exact-Match Retrieval (14.5.1), and Partial-Match Retrieval (14.5.2). We see that while the former is used for more succint/direct matches like bibliography/author name, the latter is to look at topic relevance in a text body, and uses concepts like term frequency and inverse document frequency to score the relevance of the document without being affected by stop words. We are then introduced to existing online medical information systems like Pubmed which use information retrieval techniques to provide search results to users querying for a specific medical topic.

The final topic of the chapter is about evaluation (14.6) The author first talks about the kinds of questions users might ask to determine if the system was good at information retrieval, and then talks about evaluating the system based on system-oriented evaluation (14.6.1) or user-oriented evaluation (14.6.2). In system oriented evaluation, we are reintroduced to familiar metrics like precision and recall which are covered in the course, and in user-oriented evaluation, we build on top of these metrics to find user-topic agreement and usefulness.

Finally the author talks about research directions and conclusion, we are left with a sense of great possibility in the line of using text retrieval methods in healthcare to further derive analytics to improve the patient experience. The author also points out various kinds of retrieval, and questions about what we want to retrieve/ how it helps the patients and clinicians.

Conclusion:
In conclusion, I personally think that we are at the infancy of leveraging NLP, text retrieval, and text curation techniques to vastly improve the current healthcare experience in the United States. It is a desperate problem, that needs to be solved soon, as the rates of diseases are increasing in the United States, and with more patient data to work with, we need more efficient process to manage/ retrieve this data when a patient or healthcare provider queries for it. We need processes to help Healthcare providers retrieve a transparent, linear history of the patient data so that they may be able to better understand patient history and provide adequate care for the patient. This article gave a good overview about the current standard of informational retrieval techniques in healthcare, the questions to ask while retrieving medical information and its user usefulness, and the direction in which this research will go in the future. I personally believe that text retrieval is the first step and that we need to develop robust ways to procure relevant information for healthcare providers and users. This goes beyond text retrieval to data curation, NLP, and creating user-interfaces for clients and healthcare providers alike. We are already seeing a vast amount of startups emerge in the field of healthcare, and many of them are in the business of curating medical information, and extracting relevant

patient data to provide to insurance providers and the like. I hope to play a part in this industry after completing my masters and am looking forward to the possibility of applying what I learned in this course for the same.

Reference/Citation :

Hersh W, Information Retrieval for Healthcare, in Reddy CK, Aggarwal CC (eds.), Healthcare Data Analytics, Boca Raton, FL: CRC Press, 2015, 467-505.

Thank you!