International Journal of Biomedical Data Mining

Zhu, Int J Biomed Data Min 2018, 7:2
DOI: 10.4172/2090-4924.1000135

Research Article Open Access

A Method for Searching and Ranking Medical Doctors Based on Biographical Attributes

Melinda Zhu*

Santa Monica, CA, United States

Abstract

We describe a method for constructing a searchable database for medical doctors using information extracted from unstructured natural language texts on public websites. Specifically, we focus on biographical attributes such as the doctor's medical school, undergraduate college and degree, age, medical specialties, publications on certain types of conditions (and their citation frequencies), associated media reports, etc. Ranking information for medical schools based on average MCAT scores and GPAs can be used as search parameters to provide for ranking of the search results. Citations for research publications and how often the doctor's name is associated with particular types of medical conditions can also be used for ranking purposes. Absent of any personal knowledge of a particular doctor's treatment outcomes, a patient looking for quality care can best be aided by ranked list of potential providers based on their educational backgrounds, experiences and knowledge of their specialties. Since we must collect the majority of our information from the Internet, which consists mostly of unstructured HTML based texts, finding specific information and categorizing them in a database requires natural language-based pattern recognition algorithms that can be learnt and associated with certain medical terms, as well as extract information on educational backgrounds and professional experiences. We argue that similar ideas can be applied to many other search tasks that can benefit from categorized databases built from the universe of unstructured web pages. We propose that a new type of web search engine can be designed using natural language processing techniques to mine extracted categorized information from unstructured texts to allow users to perform a variety of sophisticated searches that presently cannot be performed with current internet search engines.

Keywords: Biographical attributes; MCAT; GPA

Introduction

It is often said that patients should trust their doctors. Obviously this is overly generalizing when it comes to medical opinion. It is nevertheless an assumption that all medical doctors are trained similarly and possess the same level of medical knowledge. After all, most doctors are 'board certified' and medical degrees are awarded by a relatively small number of institutions in the United States. To become a licensed and practicing physician, it's an arduous journey of numerous checks and controls along the way to make sure that the candidate has the adequate knowledge and experience to enter a real medical profession. It can be argued that there is no other profession in the United States that goes through as rigorous a selection, training and learning process as the medical field. In reality, however, not all doctors are created equal. An area of active research is patient outcomes, especially differences in the quality and delivery of care [1-4] For many patients, when it comes to life and death decisions on which treatment plans to seek for a critical illness, they often seek 2nd or 3rd opinions from different doctors, who all have presumably undergone similar trainings for the same type of medical conditions and have studied the most recent academic literatures on the illness. However, in most cases, the decision to seek a 2nd opinion is mostly based on random factors such as proximity, or the insurance company's listing for a particular specialty. A more rational approach would be to examine specific doctors' patient outcome results for specific doctors in their areas of expertise. However, current efforts to gather Patient-Reported Outcomes (PRO) are mostly intended for medical research regarding the effectiveness of treatment options and their side effects [5]. Such data are not meant to be used by patients seeking doctors that have been successfully treating patients with certain conditions.

But not all physicians are created equal, and different doctors can have different interpretations of the PRO data as well as related medical literatures (which can be viewed as clinical reported outcomes). Some have proposed a different type of PRO research based on the physician's medical education. In "A Call for Outcomes Research in Medical Education" [3], the authors cited a study by The Commonwealth Fund Task Force on Academic Health Centers, which found that the quality of care that the public receives is determined to some extent by the quality of medical education students and residents received. It is not surprising that very few studies have been conducted on outcomes based on the medical education that the healthcare providers received. Otherwise, there would be unintended consequence of ranking doctors solely based on which medical schools they went to. But for most patients facing life and death situations, without being an expert on the treatment options for their conditions by studying the relevant medical research publications, many of them would prefer to see doctors who went to a better medical school, have been practicing for a while, and have demonstrated expertise in their field by publishing well-cited clinical studies. Granted, for many patients there is the issue of convenience in terms of the distance one has to travel to see a particular specialist, and most do not have the financial means nor the time to travel thousands of miles to another part of the country to see a well-known specialist. But the issue of convenience should be the choice of the patients, and for many who are facing serious conditions, some are willing travel the distance to seek the best care available. A system that allows patients to select doctors based on their biographical

***Corresponding author:** Melinda Zhu, Santa Monica, CA, United States, Tel: 3108669935, E-mail: zhumelinda@gmail.com

Received August 29, 2018; **Accepted** September 25, 2018; **Published** October 01, 2018

Citation: Zhu M (2018) A Method for Searching and Ranking Medical Doctors Based on Biographical Attributes. Int J Biomed Data Min 7: 135. doi: 10.4172/2090-4924.1000135

Copyright: © 2018 Zhu M. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

information can help patients make decisions regarding how to seek 2nd or 3rd opinions aside from the traditional referrals based approach [6,7].

According to a study by Kaiser Health [8,9] an estimated 10 - 20 percent of cases are misdiagnosed; 28 percent of those mistakes were life threatening or had resulted in death or permanent disability. Another study [10] on the frequency of diagnostic errors in outpatient care concluded that approximately 12 million patients in the United States are misdiagnosed each year. Published stories such as this [11] and this [12] do not inspire our trust in doctors' abilities to provide the most informed diagnoses. In the former case, a cancer patient was successfully treated only after her sister discovered a clinical trial at Johns Hopkins on her own. In the second case, a highly educated venture capitalist lost trust in the opinions of several doctors who treated him ("My experience as a patient was insane. I couldn't believe how screwed up the health system was") and took matters into his own hands, which led to an eventual cure for his rare form of cancer. Obviously, the vast majority of patients cannot rely on luck, determination (as in the sister case), or his/her own research prowess (as in the VC case) to help them make proper decisions. They would have no choice but to rely on their doctors' opinions. However, knowing there is a relatively high probability of being misdiagnosed, one would assume randomly looking up a specialist's name or going by your primary physician's recommendation would not likely lower that probability, as the majority of the patients seek care using that approach. Let's examine several factors below that might increase patients' odds of finding more informed doctors.

Methods

Medical school educations

For the majority of recent history, US medical schools rely mostly on a merit-based system to select who will be admitted into their program. To be admitted to medical schools, one has to take at least two years of chemistry, one year of biology, and one year of physics in addition to math and English requirements (more recently, students are encouraged to take additional sociology or psychology courses). These requirements are rooted in a 1910 report [13] by physician Abraham Flexner, where he stated "how much education or intelligence it requires to establish a reasonable presumption of fitness to undertake the study of medicine under present circumstances" was a "competent knowledge of chemistry, biology and physics" [14]. In 1917, Harvard professor Frederick Hammett stated, "The true physician must be a true diagnostician. He cannot be a diagnostician if he lacks the power of observation and ability to carry on deductive reasoning. Where better can he gain this fundamental training than in chemistry?" [14,15]. In order to gain admission to medical schools in the United States, students must be able to prove that they can excel in a rigorous science curriculum and demonstrate solid logical and quantitative reasoning skills. This is why almost all medical schools require their applicants to take the 7+ hours long Medical College Admission Test [16]. Despite recent efforts to admit more humanity and social science focused students, as well as broaden the requirements for extracurricular activities [17], medical schools in the US still mostly practice a merit-based admission policy based on applicant's MCAT scores and GPA (both the science and the overall part). This is demonstrated by the relatively high MCAT scores and average GPAs for the admitted students overall [18,19]. But there is a wide dispersion among medical schools in terms of matriculated students' MCAT scores and their GPAs. For example, schools ranked in the top-10 (as published by US News and World Report [20] typically have incoming students with MCAT scores in the 92-97 percentile range (as in the case of Johns Hopkins [21], Duke [22] , and University of Pennsylvania [23,24], while lower ranked schools typically have average MCAT scores in the 65-70 percentile range (as in the case of West Virginia University [25], and Florida State University [26]. Keep in mind, however, that the overall medical school acceptance rate [18], as defined by the percentage of students accepted into at least one medical school is around 32%. It is reasonable to assume that students matriculating into the lower ranked medical schools have lower MCAT scores, while the top schools tend to garner the best performing students. Given the complexity of medical diagnoses for conditions that are difficult to treat, the relatively high incidence of misdiagnosis, preliminary research linking patient outcomes to the quality of medical school education [3,7] and available quantitative academic profiles of medical students at most US medical schools, a reasonable approach is to favor doctors who demonstrate high aptitudes for analytical and quantitative reasoning, or in other words, those who graduated from the most selective medical schools. Given the opportunity, without any detailed knowledge of, or extensive recommendations from multiple sources of the medical community for a particular doctor to oversee a patient's diagnosis, one would rather select a doctor who is at the 99 percentile level among the pool of perspective medical students, than one who is at the 70 percentile level.

Misdiagnoses are often associated with unnecessarily costly treatment plans. Here, the cost can be both financially or medically detrimental, especially in the case of irreversible surgical procedures. For example, there are the widely reported issues of unnecessary heart procedures [27-30,31] and the more recent trend of unnecessary double mastectomies [27,28]. Although it is important to note that there is a difference between misdiagnosis and medical ethics, it can be argued that the practitioners' financial motivations were partly responsible for many of the unnecessary procedures [29,30].

Medical education is expensive. Currently, cost of tuition and fees at most medical schools can approach $90,000 per year. Without any form of financial aid in the form of scholarships, the total amount of student debt after 4 years of medical school can easily approach close to half a million dollars given the pace of inflation for most tuitions. Considering that the average sale price of current existing homes [31,32] in the United States is about $250,000, student loans for attending medical school could be as high as twice that of the average mortgage (assuming 20% down payments for a typical mortgage). Not surprisingly, lesser known and lower ranked medical schools tend to graduate students with the most debt [33], while many top ranked schools, with richer endowments to pay for more generous financial support, tend to graduate students with much lower average student debt [34,35]. Just as it is incorrect to assume that all unnecessarily expensive procedures are related to financial motives of the practitioners, so too is it wrong to assume that all top ranked medical school graduates have low levels of student debt. However, given the choice of doctors, patients may favor those from top medical schools due to fears regarding overtreatment [36,37].

Undergraduate institutions and majors

Before entering medical school, the majority of students follow the so-called pre-med track during their undergraduate studies. Unlike most college majors, pre-med tracks generally do not need to be declared. Rather, most pre-med students pursue various science related majors with the primary goal of completing the curriculum that can satisfy most medical schools' admission requirements. while maintaining a relatively high GPA. The pre-med curriculum is mostly centered around the subject areas of the different sections of the MCAT- in order to do well in those classes, one must you are often competing with other pre-med students for the top grades. Medical school admissions mostly look at two parts of one's GPA: the science part of the pre-med curriculum GPA, and the overall GPA. Most pre-meds' overall GPAs are generally higher than their science GPAs, as they tend to structure their major requirements around the core pre-med curriculum so that the rest of the requirements are generally 'easier' classes to obtain higher grades. For example, the pre-med curriculum does not require one to take higher level physics or chemistry courses that require calculus and differential equations, etc, or any higher level math classes beyond the required one-year basic calculus. Comparing two otherwise identical applicants in terms of their GPAs and MCAT scores, it is reasonable to assume that medical school admissions do consider the level of difficulties of each applicant's curriculum. But when choosing a doctor, one generally does not have that information from reading the background of the doctor. However, in many cases, one might deduce that from the doctor's undergraduate major. A psychology major is probably 'easier' than biology major, a bioengineering or chemical engineering major is probably 'harder' than a biology or chemistry major, etc. Similarly, a postbaccalaureate program is an alternative pre-med track typically for students who have already finished their undergraduate degree. In most cases, the goal is purely to satisfy the minimum required AAMC curriculum. Moreover, for some students, due to reasons of convenience or perceptions of easier grading, they would complete the program at schools outside from their undergraduate institutions

Besides GPA and MCAT scores, some medical school admissions also look at the quality of the candidate's undergraduate institution [38]. For many students, the schools they decide to attend for their undergraduate studies depend on many factors, such as location, financial aid, availability of particular majors, etc., but the factor of merit does play an important role in most college admission decisions . Although not as important as the quality of medical school, the quality of the doctor's undergraduate college could be one of the factors a patient should consider.

Residency

During medical school, students are required to take the United States Medical Licensing Examinations (USMLE) before obtaining their MD degrees. Unlike MCAT takers, the majority of students eventually pass this series of exams. However, students are ranked on a curve based on their scores from the USMLE exams. Residency program directors rely heavily on this ranking information to select for their residency programs medical school graduates who will be the most likely to succeed [39,40]. Similar to our arguments for looking at the the quality of the doctor's medical school, the quality of the doctor's residency program (which can depend on the specialty) could be used as another factor to consider when selecting a doctor.

Research publications, MD/PhD

Finally, we consider to what extent the doctor has published in major medical or scientific journals. In general, most patients do not care whether their doctors have published research articles in their specialty areas. It is probably more important that their doctors are intimately familiar with the latest and the most important clinical research related to the treatment options for their condition. But without any direct personal knowledge of the medical expertise of the doctor, another option is to look at whether the doctor has published in major medical journals. Since most peer reviewed medical or scientific journal publications require the author(s) to discuss and compare their findings to studies done previously by other researchers or practitioners, one can use the fact that well cited publications are indicative of expert knowledge in a doctor's specialty.

Most medical schools in the US admit a small number of students that enroll in their dual MD/PhD programs. These programs generally require higher GPA and MCAT scores, as well as extensive research experience during the students' undergraduate years. It takes almost twice as long to graduate from an MD/PhD program compared to an MD alone. Most of these MD/PhD doctors later go into practices that are associated with major research hospitals, and are active in medical and scientific research.

In general, it is expected that locating doctors who have well cited research publications to their names can be difficult when considering proximity and appointment availability factors. The same goes for MD/PhDs as their numbers can be even smaller. But if one is determined to seek out 2nd, 3rd or even more opinions, and are not deterred by the time and distance factor, it makes sense to seek out doctors who are likely to be familiar with the latest clinical research, especially in the areas that are related to one's conditions.

Approach

In the sections above, we laid out the motivations behind searching for specific attributes of doctors when the patient is faced with having to seek out multiple opinions for certain medical conditions. In general, if one is given the name of a specific doctor, it is a relatively easy task to find most of the important background information on that doctor. But it is a much harder task if one needs to find a list of candidate doctors who fit certain criteria. Despite being a rich depository of information, the World Wide Web is still consisted of mostly unstructured textual information. An internet search engine such as Google is built around a searchable database of keywords, but for the most part, the search engine does not categorize the type of keywords in its database. In addition, current Internet search engines are not Object Oriented databases. One cannot yet search for certain attributes of an object, e.g., tell me all the high schools in the city of Los Angeles that offer AP Physics C, E&M. It is possible that such information exists in some special database listing, but the information in that database is certain to have been compiled based not on unstructured Web texts (for example, by combing through all the high school websites in Los Angeles), but rather based on information that are already structured.

In order to build a searchable database for categories of objects, one would require domain knowledge of specific objects. It is perhaps theoretically possible with sufficiently advanced Artificial Intelligence techniques for computer program to gain enough domain knowledge by processing all the Web pages on the internet using Natural Languages Processing (NLP) and unsupervised Machine Learning capabilities approaching the level of a highly educated human. But an easier and more achievable approach is to let the Domain Level Experts design the structure or the model of the objects or categories, similar to the approach used by Wikipedia and many open source projects, where a number of domain level experts are in charge of maintaining and initially designing the knowledge content or the framework of specific subject areas or frameworks. Using the examples of building searchable databases for all high schools or colleges and universities, one would need to design and specify object models for high schools and universities. Here the domain knowledge model perhaps will instruct a Web search crawler what to look for when it encounters a website that resembles a school website. Here the object model for a school could consist of departments, faculty, curriculum, research areas, admissions, or perhaps financial data such as endowment and funding levels. Ideally, when all the websites have been visited by the Web crawler, the NLP part of the software will be able to parse all the relevant parts as specified in the object model for schools and populate a searchable database.

Similarly, for the main focus of our project, we are interested in building a database for doctors that can be used to perform searches based on user defined criteria as outlined in the previous sections. The object model for a doctor can be represented graphically as shown in Figure 1, where the key background attributes of a doctor are represented in four categories. The first is the quality of the medical school the doctor attended, where the quality can be defined either as some weighted combination of the admitted students' MCAT scores and GPA as reported by AAMC for all the medical schools in the US [41], or based on reputation ranking from websites such as US News [20]. We also want to see if the doctor also has a PhD, and whether the graduate school differs from the medical school (and if so, the reputation/quality of the PhD program). Next, we need to identify the name of the teaching hospital in which the doctor performed his or her residency training. For patients, this is the second most important information to learn about the background of the doctor (besides the name of the medical school), as top students usually have their pick in which residency program they want to join. Not all doctors list their undergrad education, as evident in our finding that unlike younger doctors, older doctors tend not to list their undergrad education/major on their bio pages. As discussed earlier, even though undergrad education is not nearly as important as medical education, if a doctor went to a top ranked undergrad institution (or studied a demanding subject), that information can be useful. Lastly, research publications can usually be gleaned from the doctors' bio pages if they are affiliated with a teaching or research hospital, and the importance of the publications can be measured with their citation indices.

As one can see, almost all the information that we plan to extract from unstructured web texts are nouns. The branch of information extraction (IE) within natural language processing (NLP) that can recognize and categorize (mostly) nouns is Named Entity Extraction (NER) [42]. It is an active ongoing research endeavor as interest in artificial intelligence (AI) has grown in the area of information extractions as it relates to automatic processing of natural texts on the Web, as well as medical records, legal and other documents. In general, there are three main goals of NER: (1) recognize 'names', such as a person's name, a place, a product, etc. (2) attempt to categorize these names, and (3) figure out how they relate to each other. As with most AI systems, a general purpose NER system that can handle most types of documents remains a distant and elusive goal of NLP. This is mainly because of the difficulties arising from the lack of knowledge of different domains pertaining to the documents. Therefore, the current prevalent view of the NER community is to let users tailor their own special purpose NER around general purpose NER frameworks through machine learning or rule-based methods [43]. Our approach is to use one or several relatively robust general purpose open source NER parsers and adapt them to extract specifically biographical information for doctors as discussed above.

Among the open source NER packages we have tested, we found that Stanford Named Entity Recognizer [44] performed the best in terms of being able to recognize the majority of the name and entity terms appearing in a variety of webpages we fed to the software. However, as with most NER packages, they are not tailored to any specific knowledge based domain (it is certainly not able to recognize biographical information for doctors). In the following sections, we describe our rule-based method and machine learning algorithms that can be effective for extracting information related to doctors.

Before we dive into the details of named entities and entity relationships for information related to doctors, let's look at the typical output when we run the NER parser through a webpage. The following sample biography text [45-49] describes the education background and experiences of a cardiologist:

Dr. Peter Pak practices consultative cardiology with emphasis on heart failure and cardiomyopathy. After graduating from The Johns Hopkins University with Honors in Biology he earned his M.D. degree at Harvard Medical School. He completed an internal medicine residency at Massachusetts General Hospital and a cardiology fellowship at The Johns Hopkins Hospital. Dr. Pak joined Pacific Heart Institute after serving two years on the Cardiomyopathy and Cardiac Transplant Faculty at The Johns Hopkins Hospital. He is the author of many articles in highly respected, peer-reviewed cardiology journals.

Here, the named entities are tagged with underlines and labeled with squared parentheses. The generic tags are PERSON, ORGANIZATION, and LOCATION, and as one can see that with the basic setup of the NER software, there are no tags identifying medical school names, residency hospitals, etc. The following sections discuss in detail our proposed algorithms used in identifying the major components of the information related to doctors.

Doctor name

It is relatively straight forward to identify the name of the doctor. This is because the generic NER software can effectively identify names of persons in the text. In most cases, one need only look for "Dr.", "MD", or "M.D." either before or after the name. In the case that there are multiple names showing up in the same text, it is usually a directory page listing the doctors working in a medical group. In most of these situations, besides the names of the doctors, there are usually no corresponding medical school names appearing on the same page. However, in the rare case that, e.g., we have two names with "MD" or "Dr.", but only one medical school name, we can still use this page as a cross check for a potential candidate medical school name for both doctors, as redundancy is an important part for machine learning.

Medical school name

As indicated, a generic NER algorithm can only identify ORGANIZATION names (in most cases). It does not distinguish a medical school name from a hospital name, e.g.. The most straight forward approach here is to employ a lookup database of medical school names. This is part of the commonly referred rule-based modeling in NLP. This technique should correctly identify most of the medical schools in the United States. There are several complications, however. First, a medical school can have different aliases, eg University of Chicago Medical School, Pritzker School of Medicine, or simply, University of Chicago. The lookup database would need to be as complete as possible to incorporate the most common medical school aliases. Similar issues arise with a typical internet search engine when aliases need to be correctly identified and associated with each other. But in those cases, a manually constructed lookup database is not practical for the variety and amount of data one must deal with. One would often need to use machine learning techniques in NLP as such frequency of association to identify aliases, which is possible as a typical internet search engine can be trained on an extremely large number of web pages. For our case however, it is impractical to train the alias database as there is not a large enough sample of web pages that contain different aliases of the medical schools. Luckily this task can be handled relatively easily for our case as there are only about 150 medical schools in the United States and aliases are typically only used for the well-known schools Secondly, it is often the case that two (or more) medical schools can appear in the same page. One for the school where the doctor got his/her MD, and one (or several) for residency training. In most cases, the residency name often has the word "hospital" in it, or the medical school name contains the words "Medical School". But there are cases where ambiguities can occur, such as: "Dr. Mack obtained his M.D. from Johns Hopkins and completed his oncology residency at UCSF". The simple lookup technique is not sufficient in this case. There are two approaches one can consider for this situation. The simplest one is to rely on redundancy and look elsewhere where the medical school name is explicitly spelled out as "Johns Hopkins Medical School". This is usually sufficient to resolve the ambiguity as medical school information for a particular doctor is often mentioned in multiple web pages or websites. The second approach is to use machine learning to identify medical school names when they are associated with certain key words and phrases. This is more complicated than the simple lookup approach (which works in vast majority of the situations), but it helps to solve the problem that arises when the doctor graduated from a foreign medical school, which we describe below.

In the last case, when the doctor did not obtain his/her medical degree from a US medical school , it is generally harder to gauge the quality of the medical school. Although this is not the focus of our application, the method of identifying the name of the medical school can help improve our approach to extract key information when simple database lookup fails. This involves training our NLP algorithms on a set of web pages in which the medical school names have already been correctly labeled. In particular, we look at words or phrases immediately preceding or immediately after the medical school name, and examine their frequency of occurrences. The approach can be categorized as a combination of rule-based modeling and statistical modeling in NLP. When the rules or patterns are relatively simple as in the case of medical school name, this relatively straight forward approach can work in most cases. The alternative is to apply a generalized machine learning algorithm such as artificial neural network (ANN), where we only need to feed the ANN the words around and positions relative to the medical school name. The advantages of the rule-based and statistical methods are that they generally need a smaller training set than that for ANN. Additionally, each time one fails to find the medical school for a particular doctor due to some previously unseen association, a new rule can be added or statistics can be updated, while typical ANN training would need a far greater number of examples .

Name of residency hospital/program

For the residency hospital name, one would expect to employ a similar lookup method as for the name of the medical school. But in this case, unlike for the majority of medical schools, the NER software often fails to correctly identify hospital names. As in this: completed his medical residency and cardiology fellowships there at Brigham and Women's Hospital.

and this: I then had the privilege of completing my pediatric residency at Children's Hospital Los Angeles

Clearly, the NER software sometimes does a poor job at identifying entities if popular location names are part of the entities. Unlike the strategy for medical school names, which relies on NER's ORGANIATION tags in almost all cases, we can adopt a hybrid approach as follows:

α) Match residency name database to NER ORGANINZATION tags, if any

β) Search and match regular expression (regexp) of residency names

χ) Apply rule-based methods to distinguish between residency names and medical school names

Note that regardless of whether the doctor obtained his/her medical degree from a US school or a foreign school, he/she must complete a residency at a US hospital, and a doctor can complete more than one residency program.

Undergraduate college name and major

Unlike medical school education information, not all doctors list their undergraduate education. For those that do, undergraduate education information is usually described in mostly natural language free formatted texts, e.g.:

After college at Harvard and medical school at UC San Diego, Dr. Natterson completed his medical residency, as well as his cardiology and electrophysiology fellowship, at UCLA.

We see here that three school names are mentioned in the same paragraph. A simple lookup method would have a hard time determining which school is the doctor's undergraduate institution, which is his medical school, or which is his residency hospital. As described previously, one relatively simple method is to use redundancies in other parts on the internet to figure out the doctor's medical school and residency hospital, then use method of elimination to decide which school is the doctor's undergraduate institution. Alternatively, the machine learning method described previously can also be used here, but would generally require more sample webpages to train; however, the advantage is that it does not require one to look elsewhere to find redundant information.

In principle, one can use the lookup method to extract undergraduate major information from the webpage, if present. But since majors can be described differently (depending on schools), it is better to use the machine learning method to extract information related to the doctor's major (if available). Alternatively, a rule-based method can also work effectively as the variations in sentence structure are relatively simple for undergraduate majors, e.g.: Dr. Bosserman received her undergraduate degree in biochemistry with highest honors from the University of California, Berkeley, and her medical degree from Stanford University.

Where we could look for keywords such as earned, graduated with, received, major, Bachelor, B.S. etc

Research Publications

If the doctor engages in research on a regular basis, the list of publications can often be found on the webpages that also describe the biographical background of the doctor. What we are also interested in is the citation index for the publications, which can be obtained from several sources [50,51]. Presenting the user with the highest or the average citation index of the doctor's top publications isn't particularly useful unless the user is familiar with the significance of citation index for certain areas of research. What is more useful is a percentile ranking of the doctor's publications among the doctors that engage in similar research. This requires classifying the type of research the doctor is involved in, which is unrealistic as most users lack specialized medical training or have research background themselves. Assuming we eventually can establish a database of doctors that engage in medical research, we can establish the cross citations between research publications. Our classification problem becomes similar to the clustering [52] problem in artificial intelligence. Specifically, if one thinks of links as citations or references in research publications, then our classification or clustering problem becomes similar to the clustering problem in a social network [53].

Our approach is to think of publications that are often cited as occupying the center of the cluster, and the less cited papers as more distant to the center. In this configuration, one can introduce the concept of radius, and define clusters with certain cutoff radii. Publications within each cluster or classification can therefore be ranked using their percentile rankings as defined by their citation indices.

Conclusion

Figures 1 and 2 summarize our approach to extracting biographical information for medical doctors from unstructured natural language texts on the web. The first step is to establish an object model for the background information of a doctor. The next step is to use natural language processing techniques to systematically extract key attributes of this object model. Similar to the traditional internet search engine, it is necessary to crawl through virtually the entire internet space so that a searchable biographical database can be constructed for doctors in the US.

We have argued that the quality of doctors can sometimes be the difference between life or death for many patients seeking critical care or diagnosis of potentially serious conditions. However, most patients lack the necessary medical training or knowledge to evaluate the background of their doctor. Given the lack of medical knowledge, we argue that the best approach to seeking quality care is to look for doctors who have the best medical training as the cost of care typically does not vary much from physician to physician. Without a priori knowledge of the doctor a patient is going to see, it makes sense to start by visiting doctors who graduate from the best medical schools, have widely cited publications in the field, or both. But current internet search engines only return the names of doctors, typically near where one lives. It's up to patients to spend the time going through all the web links hoping to find the relevant biographical information for the doctors and make a list to compare the results. For people seeking diagnosis or treatment of potentially critical illness, it might make sense to travel the distance to seek out experts for second, third or even fourth opinions. But web search in this case can be even more time consuming as one typically would need cover most major metropolitan areas in the US. This exposes a fundamental deficiency in current internet search engines. Given the unstructured nature of the majority of web pages, search on the internet is mostly restricted to using key words. It is not possible to search for attributes, or have the search engine return ranked results based on the attributes, even if you are allowed to define how to rank the results.

If we can systematically extract the relevant key attributes for the doctors from the web pages describing the background information for the majority the physicians that have internet presence on the US, one can build a structured database that allows users to search based on attributes and/or define how they want the results to be ranked. In this study, we presented a model about some quantifiable attributes of a doctor, and outlined an approach to extract these attributes from unstructured web texts using a combination of rule-based and machine learning natural language processing (NLP) techniques. Our main objective is not to argue that education or professional background alone should be the dominant factor in choosing a doctor, but rather a useful metric, that, if given the choice, one can make better informed choices when making healthcare decisions.

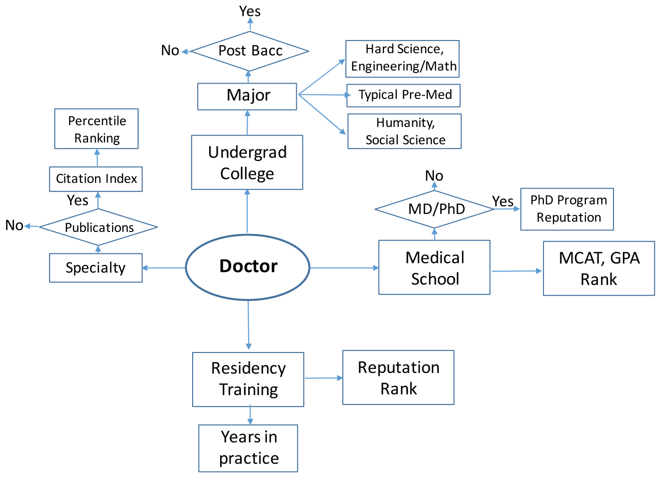

Figure 1: The Object Model of a Doctor.

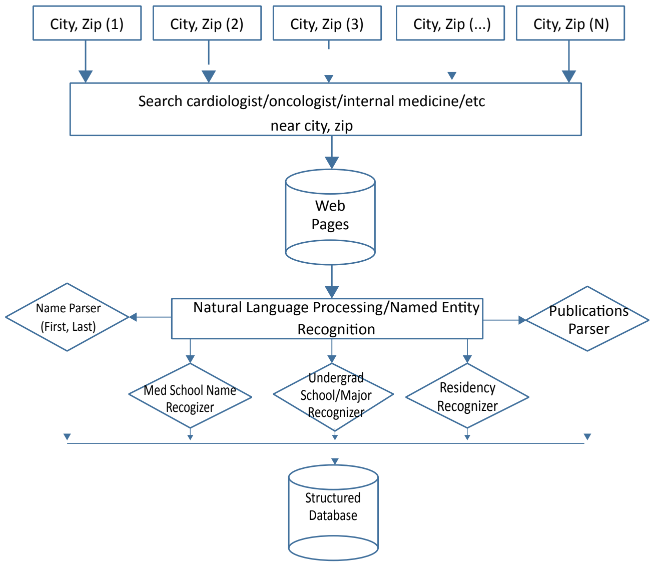
City, Zip (1)
City, Zip (2)
City, Zip (3)
City, Zip (...)
City, Zip (N)
Search cardiologist/oncologist/internal medicine/etc near city, zip
Web Pages
Name Parser (First, Last)
Natural Language Processing/Named Entity Recognition
Publications Parser
Med School Name Recogizer
Undergrad School/Major Recognizer
Residency Recognizer
Structured Database

Figure 2: Summary flow chart for extracting biographical information for doctors.

Current internet search engines essentially return ranked lists of web pages based on the search terms, which consist of series of keywords. It is the job of the users to read through these web pages to find the relevant information they need. In this setup, it is impractical to have the user go through hundreds, or thousands of web pages to tabulate the key information he/she is looking for. Since the relevancy of the search results quickly diminishes after a few pages, the more practical way would be to modify your search terms and repeat the whole exercise again, until you can build a large enough table of key information to allow you to rank what's important for you. Imagine a web crawler that does that for you already, much like how the current internet search engine was first conceived some 20+ years ago [2], but this web crawler has some basic natural language processing capabilities to extract key information for certain categories of objects. The result is a structured database that allows the user to skip the impossible tabulation process and be able to rank or filter information according to the user's own preference.

References

1. Basch, E (2017) The rise of patient-reported outcomes in oncology. American Society of Clinical Oncology. Annual Meeting, Chicago.
2. Brin S, Page L (1999) The anatomy of a large-scale hypertextual web search engine.
3. Chen F, Bauchner H, Burstin H (2004) A call for outcomes research in medical education. Acad Med 79:955-960.
4. Hess Li, Pohl G (2013) Perspectives of quality care in cancer treatment: A Review of the Literature. Am Health Drug Benefits 6: 321-329.
5. Jensen R, Snyder CF, Abernethy AP, Basch E, Potosky AL, et al. (2014) Review of electronic patient-reported outcomes systems used in cancer clinical care. J Oncol Prac10:215-22.
6. Robeznieks A (2013). AMA saw membership rise 3.2% in 2012. Modern Healthcare.
7. Training Tomorrow's Doctors: The Medical Education Mission of Academic Health Centers. The Common wealth Fund Task Force on Academic Health Centers, New York: The Common Wealth Fund 2002.
8. Physicians Misdiagnose at an Alarming Rate. National Center for Policy Analysis. 2013.
9. Boodman S (2013) Doctors' Diagnostic Errors Are Not Often Mentioned But Can Take a Serious Toll, Kaiser Health News.
10. Singh H, Meyer A, Thomas EJ (2014) The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ Qual Saf 23:727-731.
11. McGinley L (2017) This is not the end: Using immunotherapy and a genetic glitch to give cancer patients hope. Health and Sci.
12. Farr C (2017) This VC had a near-death experience and it totally changed what he invests in. CNBC
13. Flexner A (1910) Medical Education in the United States and Canada. Carnegie Foundation.
14. Lam V (2017) The pre-med drop out. The Stanford Daily.
15. Hammett F (1917) Pre-Medical training in chemistry. 46: 1195, 504-506.
16. Medical School Admission Test.
17. Muller D (2013) Reforming premedical education-out with the old, in with the new. J Med 368:1567-1569.
18. MCAT and GPA Grid for Applicants and Acceptees to US Medical Schools. Association of American Medical Colleges (AAMC), 2018.
19. MCAT Scores and GPAs for Applicants and Matriculants to U.S. Medical Schools. Association of American Medical Colleges (AAMC), 2018.
20. Best Medical Schools, US News and World Report, 2018.
21. Class Statistics-Johns Hopkins School of Medicine, Johns Hopkins University, 2018.
22. Facts and Figures-Duke University School of Medicine, Duke University, 2018.
23. Entering Class Profile-Perelman School of Medicine.
24. Percentile Ranks for the MCAT Exam, AAMC, 2018.
25. M.D. Student Services-West Virginia University School of Medicine, West Virginia University, 2018.
26. M.D. Admissions FAQ-FSU School of Medicine, Florida State University, 2018.
27. Wong S, Rachel A, Yasuaki S, Fatih A, Wlliam B, et al. (2017) Growing use of contralateral prophylactic mastectomy despite no improvement in long-term survival for invasive breast cancer. Ann of Surg 581-589.
28. Sternberg S, Dougherty G (2015) Are doctors exposing patients to unnecessary cardiac procedures. U.S. News & World Report.
29. Kavarana M, Sade R (2012) Ethical issues in cardiac surgery. Future Cardiol 8:3.
30. Charatan, F (2003) Dozens of patients allege unnecessary heart surgery. British med J.
31. Yao, L (2009) Common heart surgery may be unnecessary. Wall St J.
32. Median Sales Price of Existing Homes. Federal Reserve Bank of St. Louis, 2018.
33. Which medical school graduates have the most debt ? U.S. News and World Report 2018.
34. Facts and Figures-Harvard Medical School. Harvard University Medical School 2018.
35. Scholarships-Duke University School of Medicine, Duke University, 2018.
36. Chen, P (2013) The changing face of medical school admissions. New York Times.
37. Martin, S (2008) Stanford pre-meds spend summer at santa clara university, where physics is easier. The Mercury News.
38. Requirements-University of Michigan Medical School. University of Michigan, 2018.
39. Results of the 2016 NRMP Program Director Survey, National Resident Matching Program, 2016.
40. Charting Outcomes in the Match, National Resident Matching Program, 2014.
41. Medical School Admission Requirements (MSAR), AAMC, 2018.
42. Information Extraction and Named Entity Recognition", Stanford University, 2018.
43. Nadeau D, Sekine S (2007) A survey of named entity recognition and classification. Lingvisticæ Investigationes 30:1.
44. Stanford Named Entity Recognizer, Stanford University, 2018.
45. Open NLP, Aparche.org.
46. General Architecture for Text Engineering.
47. Named Entity Recognition and Disambiguation.
48. Spacy.
49. Pacific Heart Institute.
50. Active Physicians Who are International Medical Graduates (IMGs) by Specialty, 2015.
51. Citation Index-Wikipedia.
52. Carnegie Mellon University.
53. Mishra N, Clustering Social Networks, Stanford University.