Towards Automated Healthcare for the Masses Using Text Mining Analytics

Vishnu S. Pendyala

Abstract— Current health system cannot sustain due to its demand for increasing finance and human capital resources. This article is an attempt to profess the need for deploying text mining analytics for healthcare, so as to make it available to large populations of the world. The article presents a brief survey of the work towards the goal and points to two other papers written by the author on the topic for those interested in knowing the specifics.

Index Terms—Text Mining, Healthcare, Expert System, Information Retrieval, TF-IDF

1 Introduction

THE recently released National Health Policy 2017 by Government of India proposes to assure healthcare for all Indian citizens, falling just short of making healthcare a fundamental right. It is a move in the right direction, but it still seems an ambitious goal. Even when USA spent more than 17% of its GDP on healthcare costs (Source: http://worldbank.org), it could not assure healthcare as a fundamental right to its citizens. India's National Health Policy proposes a time-bound increase to a mere 2.5% of the GDP for healthcare.

It is a fact that a vast portion of the world population does not even have access to proper healthcare and the cost of healthcare is steeply increasing. Both are serious concerns that need to be addressed. If feasible, automated diagnosis, which is machines doing the diagnosis instead of human doctors, will substantially help in both respects. The constitution of World Health Organization considers "highest attainable standard of health" as a fundamental right of the people. How governments provision this right is left to the nations and sadly, many governments are not in a position to pass on this right to their citizens. Many governments cannot afford to provide the right and access to even basic healthcare is not at an acceptable level in many regions of the world.

2 International Efforts

Pandemics and disasters will only increase with time because people travel. But more than pandemics and Communicable Diseases, as populations prosper, it is the Non-Communicable Diseases (NCD) such as diabetes and cardio vascular disease that become more life threatening. The two apex bodies in their respective fields, The International Telecommunication Union (ITU) and World Health Organization launched a four-year initiative to use mobile technologies to combat NCDs such as hypertension and cancer.

The project has demonstrated improvements in a) Disease and Epidemic Outbreak Tracking b) Mobile Health care telephone help line, c) Treatment compliance, d) Appointment reminders, e) Community mobilization, f) Mobile surveys (surveys by mobile phone), g) Surveillance, h) Patient monitoring, i) Information and decision support systems, Pregnancy care advice by SMS j) Patient record keeping. However, it still leaves much to be done. With the increasing finance and human resources capital, it is almost impossible for the current healthcare system to scale to the future need. We therefore need an automated way of supplementing the healthcare system. We want to invent drastic solutions to help the masses.

3 Technological Approaches

An analysis of the past and present trends show that Computing Technology has been advancing at an exponential rate from the times known to the mankind. The human genome sequencing project, which costed more than a bil-

lion dollars a decade ago (Source: http://www.genome.gov), is now targeted to be done in just a thousand US Dollars. A substantial number of tasks that only humans could do, are now completely automated, resulting in a far better quality of life.

In spite of all this progress, one function that has remained an unsuccessful target for automation for decades is the human expertise. While other branches of Artificial Intelligence (AI) have progressed in leaps and bounds and have come a long way in all these decades, the Expert Systems branch of AI that models human expertise has not really taken off even after so many years. Expert Systems used for medical diagnosis, such as MYCIN developed almost four decades ago, even that day, with the limited infrastructure, actually outperformed human experts.

Still, in spite of the huge gains in infrastructure and computing speeds, the idea of automated medical diagnosis for mass and general application is not widely used for various reasons, some of which are beyond the scope of this paper. It may still be worth revisiting that idea again, but with an entirely different approach, with the hope that one day, automated general medical diagnosis not specific to any disease or condition, becomes a reality.

4 Problem Reoriented

Earlier Expert Systems were rule based. Knowledge was modeled and reasoned using first-order-logic systems. A number of rules were put in place so that the system could reason with them. The end result was still Information Retrieval (IR) - retrieving the information needed by the user. We can therefore reorient the problem primarily as that of Information Retrieval and not as a knowledge engineering or expert reasoning problem.

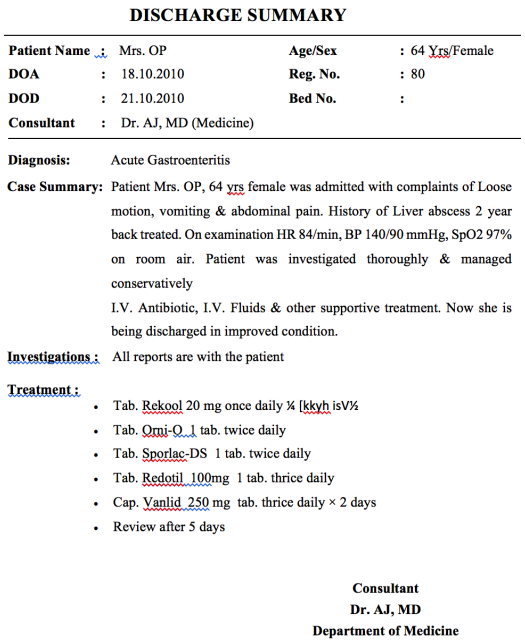
Developments in IR now make it possible to try self-diagnosis by doing an Internet search. It is quite typical to search for symptoms on the Internet to get an idea of the disease, before receiving professional help. So, it can be hypothesized that the IR as a technology is mature enough to be used for professional medical diagnosis. The conventional programming paradigm has been to generalize from small data. The current trend is to personalize from big data using IR tools. The legacy of deterministic programming is fast yielding to the new model of probabilistic programming to deal with the uncertainty in the big data.

5 Enabling Factors

A few key enablers can make this happen: a) Mobile is doing what the World Wide Web did in the nineties, which is taking technology and services to the masses. b) Machines already proved that diagnosis can be automated. c) We already have ways to cure at the molecular level by replacing mutated or damaged genes. d) There are companies which manufacture wearables that can measure the vitals and other health signals such as ECG. e) We so far could not replace the mind, but we were successful in replicating some of its functionality. All these pieces of the puzzle are combined to unfold a vision for making automated diagnosis available to the masses using mobile devices connected to the cloud infrastructure which is presented in the author's work cited as [11] and [12].

6 Methodology

The crux of the work described in these two papers, [11] and [12] can be described in a few sentences. The first step is to collect huge sets of "discharge sheets", which contain a systematic description of the symptoms and the diagnosis of a medical expert. This is the big data or "text corpus" involved in the project. A sample discharge sheet is shown in Figure 1. Each of the discharge sheet is plotted as a vector in a multi-dimensional space. Each word in the corpus of the discharge sheet is a dimension. The value along the dimension is the TF-IDF score of the document. TF-IDF stands for "Term Frequency–Inverse Document Frequency." It is a measure of a word's relative importance to a document. If a word occurs in all the documents in the corpus, its importance to any single document is low. Hence the term, "inverse" in front of the "document frequency."



DISCHARGE SUMMARY

Patient Name	: Mrs. OP	**Age/Sex**	: 64 Yrs/Female
DOA	: 18.10.2010	**Reg. No.**	: 80
DOD	: 21.10.2010	**Bed No.**	:
Consultant	: Dr. AJ, MD (Medicine)		

Diagnosis: Acute Gastroenteritis

Case Summary: Patient Mrs. OP, 64 yrs female was admitted with complaints of Loose motion, vomiting & abdominal pain. History of Liver abscess 2 year back treated. On examination HR 84/min, BP 140/90 mmHg, SpO2 97% on room air. Patient was investigated thoroughly & managed conservatively

I.V. Antibiotic, I.V. Fluids & other supportive treatment. Now she is being discharged in improved condition.

Investigations : All reports are with the patient

Treatment :

- Tab. Rekool 20 mg once daily ¼ [kkyh isV½
- Tab. Orni-O 1 tab. twice daily
- Tab. Sporlac-DS 1 tab. twice daily
- Tab. Redotil 100mg 1 tab. thrice daily
- Cap. Vanlid 250 mg tab. thrice daily × 2 days
- Review after 5 days

Consultant
Dr. AJ, MD
Department of Medicine

Figure 1 A Sample Discharge Sheet

The discharge sheets are now vectors in multidimensional space. When a new set of symptoms arrive, they can be documented and plotted just like described above using TF-IDF scores in the same multidimensional space. The only difference from the discharge sheets is that this new document of symptoms does not have the diagnosis listed on it. If we can find a discharge sheet that is similar to the document of symptoms, we can use the diagnosis listed on the similar discharge sheet for the set of given symptoms. Finding similar discharge sheet is a matter of identifying the closest vector to the vector representing the symptoms. Closest vector to a given vector can be identified by computing the cosine similarity that we learnt in high school math.

7 Improving the Accuracy by using Related Work

These ideas presented in the two papers can be combined with other related work to improve the accuracy of the diagnosis. For instance, using augmented reality enabled Web, an image of a patient's MRI can be superimposed on his body to provide further information to the diagnosis application that can be used to further confirm or fine-tune the diagnosis done using text mining.

Missing data is a common problem with Clinical repositories, including discharge sheets. Purwar et al [2] present a novel way of imputing missing values using simple K-means clustering and use it for predicting disease onset, with highly accurate results. Authors of [3] apply a sequential learning framework to model and predict the progression of Alzheimer's Disease in a patient. The dataset they used for the purpose is humongous clinical diagnosis data, particularly comprising of the medical images of brain scan. Karamanli and others [4] use Artificial Neural Networks on clinical data to predict cases of Obstructive Sleep Apnea. Constantinou et al [5] use Bayesian Networks to model expertise to support medical decisions.

As mentioned in the introduction, self-diagnosis is becoming popular and [6] examines its effectiveness by evaluating the symptom checker apps. They conclude that there are issues with the symptom checkers. The authors of [7] use K-means clustering for knowledge extraction for improving the prediction of traumatic brain injury survival rates. A layered approach to data collection, management, and providing services out of the data is presented in [8] using Cloud and Big Data Analytics. The authors of [9] present a mobile application that deploys an intelligent classifier to predict heart disease. They use machine learning algorithms on clinical data to do this. Integrated with the mobile application is a real-time monitoring component that constantly monitors the patient and raises an alarm when the vitals flag an emergency. In [10], the authors discuss organizing breast imaging examinations and mining them for structured reporting.

Some other techniques for diagnosis include using Multiple Logistic Regression (MLR) and Sequential Feature Selection (SFS) on a Coronary Artery Disease (CAD) dataset to select features as described in [1] and apply Neuro Fuzzy Classifier (NFC) for CAD diagnosis. All these techniques can be used as an ensemble approach along with the Text Mining technique that is described in the article to automate medical diagnosis and improve its accuracy.

8 Conclusion

In this paper, we listed a few ways diagnosis can be automated. Self-diagnosis by doing a Web search often results in correct identification of the problem, indicating that the information online is quite useful, when it comes to medical diagnosis. Any medical diagnosis application may be missing out significantly, if it does not leverage the humongous information available on the Web. This probably is one future direction that interested readers can explore.

References

[1] Marateb, H. R., & Goudarzi, S. (2015). A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system. Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences, 20(3), 214.

[2] Purwar, A., & Singh, S. K. (2015). Hybrid prediction model with missing value imputation for medical data. Expert Systems with Applications, 42(13), 5621-5631.

[3] Xie, Q., Wang, S., Zhu, J., Zhang, X., & Alzheimer's Disease Neuroimaging Initiative. (2016). Modeling and predicting AD progression by regression analysis of sequential clinical data. Neurocomputing, 195, 50-55.

[4] Karamanli, H., Yalcinoz, T., Yalcinoz, M. A., & Yalcinoz, T. (2016). A prediction model based on artificial neural networks for the diagnosis of obstructive sleep apnea. Sleep and Breathing, 20(2), 509-514.

[5] Constantinou, A. C., Fenton, N., Marsh, W., & Radlinski, L. (2016). From complex questionnaire and interviewing data to intelligent Bayesian network models for medical decision support. Artificial intelligence in medicine, 67, 75-93.

[6] Semigran, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2015). Evaluation of symptom checkers for self diagnosis and triage: audit study.

[7] Rodger, J. A. (2015). Discovery of medical Big Data analytics: Improving the prediction of traumatic brain injury survival rates by data mining Patient Informatics Processing Software Hybrid Hadoop Hive. Informatics in Medicine Unlocked, 1, 17-26.

[8] Zhang, Y., Qiu, M., Tsai, C. W., Hassan, M. M., & Alamri, A. (2015). Health-CPS: healthcare cyber-physical system assisted by cloud and big data.

[9] Otoom, A. F., Abdallah, E. E., Kilani, Y., Kefaye, A., & Ashour, M. (2015). Effective Diagnosis and Monitoring of Heart Disease. heart, 9(1).

[10] Margolies, L. R., Pandey, G., Horowitz, E. R., & Mendelson, D. S. (2016). Breast Imaging in the Era of Big Data: Structured Reporting and Data Mining. AJR. American journal of roentgenology, 206(2), 259.

[11] Pendyala, V. S., Fang, Y., Holliday, J., & Zalzala, A. (2014, October). A text mining approach to automated healthcare for the masses. In Global Humanitarian Technology Conference (GHTC), 2014 IEEE (pp. 28-35). IEEE.

[12] Pendyala, V. S., and Figueira, S. (2017, April). Automated Medical Diagnosis from Clinical Data. In IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), 2017 IEEE.

Vishnu S. Pendyala is a Senior Member of IEEE and Computer Society of India, with over two decades of software experience with industry leaders like Cisco, Synopsys, Informix (now IBM), and Electronics Corporation of India Limited. Vishnu received the Ramanujam memorial gold medal at State Math Olympiad and has been a successful leader during his undergrad years. He also played an active role in Computer Society of India and was the Program Secretary for its annual convention, which was attended by over 1500 delegates. Marquis Who's Who has selected Vishnu's biography for inclusion in multiple of its publications for multiple years. He is currently authoring a book on a Big Data topic to be published by Apress / Springer.