The AAAI 2017 Spring Symposium on
Artificial Intelligence for the Social Good
Technical Report SS-17-01

In Search of Health Doubles

Malay Bhattacharyya
Department of Information Technology
Indian Institute of Engineering Science and Technology, Shibpur
Howrah – 711103, India
E-mail: malaybhattacharyya@it.iiests.ac.in

Abstract

With the advent of smartwatches and smartphones, healthcare support has recently become personalized. The dynamic health related data of an individual can be easily sent to anywhere around the world for the purpose of supportive actions. In this paper, we add an extra dimension to this progress by posing a novel problem, hereafter denoted as the problem of identifying *health doubles* of a particular person. By the term health double, we symbolize the persons identical in health conditions. To be precise, we show how one can address the problem of finding persons over the globe whose current health parameters predominantly match with another person. Understandably, the health doubles of a person constitute a dynamic set, thereby making the problem more challenging. We explain through a working crowd-powered model how, with the synergy of AI and HCI, the said problem can be managed. Additionally, state-of-the-art challenges of addressing the problem of exploring health doubles are also discussed in detail.

Introduction

Development of low-cost sensors and progress in Information Technology have both paved the way of making smart devices available to majority of the people. With smartwatches and smartphones in hand, any individual can take help of the healthcare support online (Boulos et al. 2011; (Ed.) 2015). This makes the healthcare support personalized (Tibrewal, Singh, and Bhattacharyya 2016). At the same time, it opens up a new promise of research in the area of big personal health data analysis. The data being dynamic and available anywhere on the globe for supportive actions highlights some challenging areas of research. In this paper, we pose a new problem, hereafter denoted as the problem of identifying *health doubles* of a particular person. As the name suggests, this problem aims to find the set of persons anytime and anywhere in the world whose current health parameters predominantly match with a given person.

Let us explain the relevance of this problem with a simple example. Suppose in one fine morning a doctor tells someone that he has been diagnosed with a very rare disease. The doctor also admits that there is a rare chance of treatment of the patient unless the disease model is realized in a greater depth. But for this, he needs to study more such subjects affected by the same disease. A similar experience was acquired by a faculty member of University of Utah, namely Matt Might, a couple of years ago when he found that his son is affected by a very rare disorder known as N-Glycanase deficiency (Might). At this stage, there remains the only option of finding patients who have the same disease around the globe. When the decision has to be made based on a health associated dataset that comprises only your data then it becomes an almost impossible task (Estrin 2014). So, we look forward to enriching the data from a global perspective. On a lighter side, this problem has another common version that is often experienced in many countries where blood banks are not well managed. The persons having rare blood groups often look for donors (having the same blood group) in case of an emergency and this becomes a real pain. All the above examples basically demand for an organized approach for finding the health doubles of a person in a broader sense. In reality, solving such problems requires a heavy teamwork over social media.

By introducing the term health double, we would like to refer to a person (may be residing in the other half of the globe) whose current health parameters predominantly match with another person. In fact, one could possibly have a set of such heath doubles and the set is expected to be dynamic. How teamwork based approaches can lead to a better hunt for health doubles for any person is a challenging task both in HCI and AI. The search can be guided by HCI and the classification of health doubles from health related features can be done with the support of AI. This paper aims to address these issues.

The current paper is organized as follows. We first address the problem of quantifying a health double. Thereafter, we describe a crowd-powered approach that can be used for identifying health doubles dynamically and fast. Next, we describe the challenges experienced in course of this implementation and finally conclude the paper.

Quantifying a Health Double

For realizing the problem in a better way, it is important to first understand how we can define a health double. We approach this issue formally in the successive subsections.

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

A schematic view of the problem

We show a detailed schema that can be used for modeling the problem of quantifying health doubles in Fig. 1. This entire schema is based on the collection of features and their dynamic processing. The features can be categorized into the following classes.

- **Static Features:** This includes the constant features of a person. E.g., the current age, sex, ethnicity, blood group, etc. of a person.
- **Medicinal Features:** This includes the features related to the medical background of a person. E.g., the history of vaccination, medical history, etc. of a person.
- **Dynamic Features:** This includes the dynamically changing health parameters of a person. E.g., the levels of blood pressure and sugar, pulse rate, etc. of a person.

Note that, we denote "age" as a static feature simply because it has a constant predictable pattern although it changes over time. Interestingly, different classes of features have varying impact on the human health and therefore can affect the recognition of health doubles in diverse ways.

Quantifying human-human association

The strength of association $F_{i,j}(t)$ (with respect to all the features), for a time interval t, is calculated for a human from the fraction of association with a particular feature. However, it is not very simple to define such a measure of strength. Some features (especially the medicinal features) may be found to have connection with a specific human only, however, there is no guarantee that it is connected to no more human. This might happen due to the diversity of diseases or the limitation like data collection. To take care of this heterogeneity, we have used the notion of frequency-inverse document frequency (Salton, Wong, and Yang 1975) to measure the value of $F_{i,j}(t)$. For a particular human i and feature j, it is calculated as

$$F_{i,j}(t) = f_{i,j}(t) \log \frac{T_H}{N_i}, \tag{1}$$

where $f_{i,j}(t)$ represents the absolute co-occurrence between a pair of human and feature for a given time interval t, T_H denotes the number of all humans in the dataset, and N_i is the number of humans where feature i appears. Since all the features in reality have at least one associated human, the potential problem of dividing by zero does not arise.

To quantify the similarity between a pair of humans, we denote the association pattern of a particular human k with a set of features as a weight vector H_k (say feature association vector) as follows

$$H_k = [F_{k,1}(t), F_{k,2}(t), \ldots, F_{k,n}(t)], \tag{2}$$

where $F_{i,j}(t)$ denotes the strength of the association between the human i and feature j for a given time interval t. In this way, if there are N number of features then the dimension of H_k is N for all the humans.

Based on this, the cosine similarity between a pair of feature association vectors H_x and H_y of a pair of humans x and y, respectively, is defined as

$$\mathcal{S}(H_x, H_y) = \frac{\sum_i H_{x,i} * H_{y,i}}{\sqrt{(\sum_i H_{x,i}^2) * (\sum_i H_{y,i}^2)}}, \tag{3}$$

In the current analysis, the type of association that might exist between a human and a feature can be of two types, either negative (bad) or positive (good). For example, having a disease is a negative feature for a human and vaccination is a positive feature. Combining these two separate types of feature associations to measure the overall human-human similarity will be erroneous. Therefore, we have calculated the similarity between two humans separately from the negative and positive feature association information. Suppose, these similarities are denoted as $\mathcal{S}^-()$ and $\mathcal{S}^+()$ for negative and positive cases, respectively. Finally, we can derive the overall similarity between a pair of humans as the weighted average of these two similarity values as shown below

$$\mathcal{S}^*(H_x, H_y) = \frac{C^- * \mathcal{S}^-(D_x, D_y) + C^+ * \mathcal{S}^+(D_x, D_y)}{C^- + C^+}, \tag{4}$$

where C^- denotes the total number of features that have negative association with the humans x and y taken together. Similarly, C^+ denotes the total count of positive features. If a pair of humans x and y produce a high similarity score, then we infer a high association between these humans. The more the similarity, higher is the chance of being similar to each other for being considered as health doubles. Note that, $\mathcal{S}^*(H_x, H_y)$ ranges between $[0, 1]$.

Searching for Health Doubles

The entire process of searching for health doubles can be segregated into the following three principle phases of operations.

1. Knowledge extraction
2. Learning
3. Search

The extraction of features, learning from them, and finally gathering knowledge is simply the tasks of AI. We add the power of HCI with it for a preparing a working model.

Let us first introduce some terminologies that will be used hereafter. A network $N = (V, A)$ is defined with a set of nodes $V = \{v_1, v_2, \ldots, v_{|V|}\}$ and a set of arcs $A : (v_i, v_j)$ $(v_i \neq v_j, \forall v_i, v_j \in V)$, which connect these nodes. Generally, we discard self-loops or parallel arcs from a simple network and consider it to be undirected. Whenever a network is called directed, we distinguish between the two arcs (v_i, v_j) and (v_j, v_i) $(\forall v_i, v_j \in V)$. A subnetwork $N' = (V', A')$ is a part of the network $N = (V, A)$ such that $V' \subseteq V$ and $A' \subset A$. Again, by the term induced subnetwork we restrict A' to include only the comprehensive set of arcs existing within the nodes of V' in N.

To better represent the global view of the crowd-powered system, we model it as a network. The persons involved in the system are basically the nodes of the network forming a large scale connectivity. It is realizable that success of the

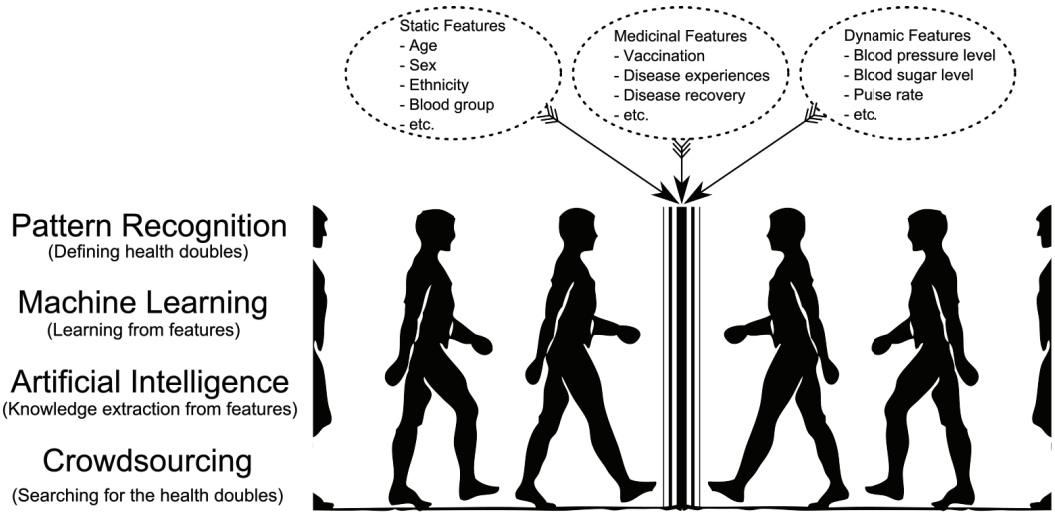

Figure 1: The schema of identifying health doubles of a particular person. Different classes of features can be collected from persons and thereafter processed for finding out the closely matching sets.

search for health doubles lies in the speed of information flow through the network because the data is continuously changing. We can think about a naive approach where all the people communicate with each other in pairs for propagating the dynamic health-related data. But unfortunately, this will involve a communication complexity of $\binom{m}{2} = O(m^2)$ for m number of people.

Crowdsourcing has recently become a useful tool for rapid processing of information for diverse purposes (Kittur et al. 2013). We introduce a crowd-powered approach that can be useful for speeding up the mechanism of identifying health doubles. It basically uses crowd as the medium of hierarchical information flow for collecting data from the crowd. We assume a network of crowd and propagate the data by employing an efficient mechanism to make the overall process parallel and dynamic. This mechanism is highlighted in Fig. 2. Understandably, the communication complexity reduces to $O(\log m)$ with such a parallelized mechanism.

Challenges

The major challenge, as we have experienced in course of the current study, behind identifying health doubles is basically defining the health doubles in an appropriate manner, given the dynamic set of parameters of a person. It is understandable that although it is very challenging to explore the exact health doubles, however it is possible to obtain an approximate set. Again, it is really impossible to find health doubles with a higher score of match (as explained earlier) because of the diversity of parameters and health conditions of different people. So, we had to take a very low threshold of similarity score. Finally, we realize that the success of the proposed model entirely depends on the involvement of crowd. More the inputs from crowd, more the data, more its reachability around the world, thereby ensuring a successful search for health doubles. Given these requirements, it might

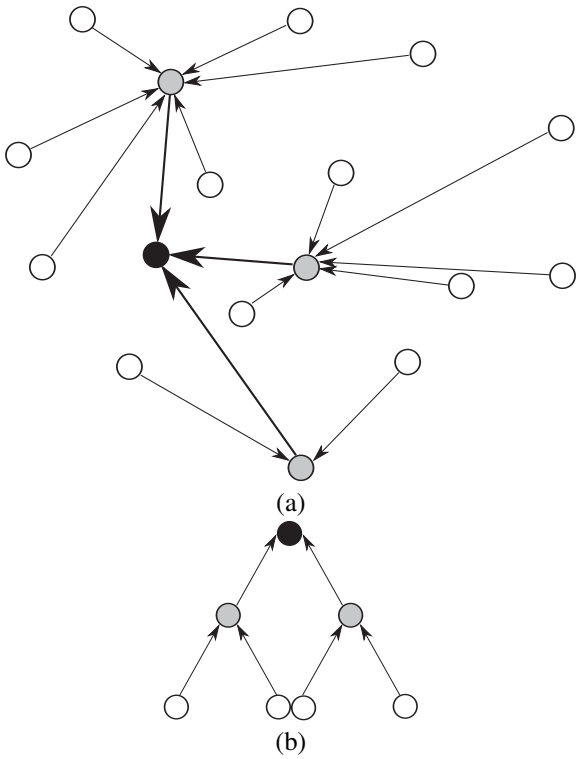

Figure 2: (a) A crowd-powered dynamic approach for identifying health doubles. The darker nodes are collecting data from the lighter nodes through a network of crowd. (b) The complexity of hierarchical data collection from the crowd.

be helpful to combine such models with the power of social networks for a better performance.

Conclusion

In this paper, we highlight that how an appropriate synergy of AI and HCI can help to address the problem of identifying health doubles over the globe. We present a crowd-powered model for addressing this problem. It shows a high promise although there exist some challenges that are still to be addressed for a fully functional model.

Acknowledgments

The work of Malay Bhattacharyya is supported by the Visvesvaraya Young Faculty Research Fellowship 2015-16 of DeitY, Government of India.

References

Boulos, M. N. K.; Wheeler, S.; Tavares, C.; and Jones, R. 2011. How smartphones are changing the face of mobile and participatory healthcare: an overview, with example from eCAALYX. *Biomedical Engineering OnLine* 10(1):1.

(Ed.), S. A. 2015. *Mobile Health: A Technology Road Map*, volume 978-3-319-12817-7. Springer.

Estrin, D. 2014. Small data, where n = me. *Communications of the ACM* 57(4):32–34.

Kittur, A.; Nickerson, J. V.; Bernstein, M. S.; Gerber, E. M.; Shaw, A.; Zimmerman, J.; Lease, M.; and Horton, J. J. 2013. The Future of Crowd Work. In *Proc. CSCW 2013*, 1301–1318. San Antonio, Texas, USA: ACM Press.

Might, M. Hunting down my son's killer.

Salton, G.; Wong, A.; and Yang, C. S. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18:613–620.

Tibrewal, R.; Singh, A.; and Bhattacharyya, M. 2016. mSTROKE: A Crowd-powered Mobility towards Stroke Recognition. In *Proceedings of the* 18^{th} *International Conference on Human-Computer Interaction with Mobile Devices and Services*, 645–650.