Predicting Discrete Reason for Visit using Personalized PageRank

John Schrom, MPH; Emma Sahn; Bobby Caplin, MBA; Vijan Joshi, MD
One Medical, San Francisco, CA

Introduction

Patients generally provide a reason for visit when booking a primary care appointment. This reason is useful for provisioning care, monitoring utilization, forecasting future patient and clinic needs, and assessing risk. There are a finite set of reasons patients seek primary care, and many are often readily apparent (e.g., follow-up from a recent hospitalization). Tethered personal health records increasingly allow patients to book their own appointments online, creating opportunities for improved user experience and enhanced data collection. This preliminary work seeks to understand the feasibility of predicting a patient's reason for visit prior to appointment booking, in order to help improve the user experience of booking appointments via a personal health record.

Methods

Data Collection and Processing. Data was collected from the 2014 and 2015 National Ambulatory Medical Care Survey (NAMCS), including discrete reason for visit, prevalent diagnoses, continued (i.e., not new) medications, patient demographics (age, sex, pregnancy status), and visit characteristics (time of year, previous visits with same provider). The binarized data was represented as an undirected graph, with the adjacency matrix formed by the number of shared encounters between two concepts.

PageRank. The PageRank algorithm[1] works by assigning each node in a graph some initial value (R_0). These values are then distributed throughout the network via an adjacency matrix (A). This is repeated recursively until convergence. While PageRank is static and deterministic for a given graph, this process can be further personalized[2]. A personalization vector, v, can be added to the equation with some weight (alpha).

$$R_{t+1} = \alpha v + (1 - \alpha) A R_t$$

In this implementation, all non-"reason for visit" nodes are included in the personalization vector, while only the "reason for visit" nodes in R are examined post-convergence.

Evaluation. The adjacency matrix was generated using 2014 data, and evaluation was carried out on 2015 data. Hyperparameter tuning was done via grid search. Accuracy was calculated as the primary evaluation metric.

Results

In total, there were 16,846 clinical encounters included in the analysis, with 2,667 binarized variables. The average unscaled edge weight was 9.8 encounters, and the average degree was 143 (range: 8 - 2666). Two nodes were connected to more than 90% of the graph: patient having "been seen before" (degree: 2613), and patient "not having any chronic conditions" (degree: 2666). The personalized PageRank algorithm correctly identified reason for visit for 31% of evaluation cases.

Discussion

These results suggest that framing reason for visit prediction as a graph problem, and using personalized PageRank as a solution, may be an effective approach. While the results leave room for improvement, it performs well given the large search space and sparse data set. This method could be improved by pruning the graph of highly connected nodes, or adding additional data (e.g., additional years) or data types (e.g., labs). Improvements in evaluation could also be informative (e.g., probability threshold for inclusion, or top N predicted reasons).

Further study is needed to understand patient acceptance of discrete reasons for visit, and how suggesting a reason for visit might influence or change patient-reported data. However, this approach shows promise as a method of improving usability and data collection during appointment booking in a tethered personal health record.

References

1. Page L, Brin S, Motwani S, Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab, Nov 1999.
2. Haveliwala T et al. An analytical comparison of approaches to personalizing PageRank (Technical Report). Stanford, CA: Stanford University. http://ilpubs.stanford.edu:8090/596/1/2004-45.pdf