Predicting Discrete Reason for Visit using Personalized PageRank

John Schrom, MPH, FAMIA; Emma Sahn; Bobby Caplin, MBA; Vijan Joshi, MD

Introduction

When making a clinical appointment, patients are typically required to provide a reason for seeking care. This reason for visit is used by clinical and administrative staff to provide care, monitor utilization, forecast future patient and clinic needs, and assess risk/predict future outcomes.

In a primary care practice, there is a finite set of reasons patients seek care: annual physicals/ preventive care, medication management, chronic disease management, and acute illnesses. Empirically, the reason for visit is often readily apparent (e.g., follow-up from a recent appointment or hospitalization), suggesting the process of soliciting reason for visit from patients could be augmented or automated.

The idea of recommending or predicting user behavior is quite common outside the biomedical domain: companies like Google and Facebook built their entire companies around such predictions. At these companies, many large-scale problems are framed as network or graph analysis problems. This has led to the development of powerful algorithms, like PageRank, to parse through the relationships of large numbers of concepts.

These algorithms aren't completely new to medicine. There have been implementations used to solve targeted medical problems, generally revolving around improving access and comprehension of clinical data for patients[1], or improving ability of clinicians to find relevant data in unstructured electronic health record data [2,3]. The problem of predicting reason for visit shares similarities with these problems: it is a high-dimensional domain seeking predictions of many different classes.

This preliminary work seeks to explore the utility of network approaches in predicting reason for visit prior to a clinical encounter.

Methods

Data Collection and Processing

The National Center for Health Statistics within the Centers for Disease Control conducts and publishes surveys of patient care in ambulatory practices through the National Ambulatory Medical Care Survey (NAMCS). NAMCS data contains coded reason for visits, along with patient demographics and comorbidities, prescribed medications, procedures, common lab and vital values, and visit characteristics.

Data was collected from the 2014 and 2015 NAMCS datasets[4]. Records were included if they were for primary care (i.e., family practice or internal medicine), and for patients over age 15. Medications were included only if they were a continued (i.e., not new) prescription. Similarly, only prevalent diagnoses were included. Other relevant visit and patient characteristics were also included only if they would be known prior to the visit itself (e.g., patient age, sex, pregnancy status, previous visits with the same provider, time of year). The resultant dataset was binarized, with data from 2014 used for model building and analysis, and 2015 data held out for evaluation.

Data Collection and Processing

A graph, G, consists of a set of vertices, V. For this problem, V was defined to be any binarized concept in the dataset (e.g., "male sex", "age 15 - 24", "has depression", "reason for visit: hypertension").

These vertices are connected by a series of edges, E. These edges are weighted according to the function phi, defined as the number of shared encounters containing concepts in their respective vertices. The weights are further scaled by the total number of encounters containing each concept, such that the sum of all the weights for a given concept is one.

The entire graph can be represented as an adjacency matrix, A. This p x p matrix (where p is the number of vertices) consists of the edge weight between the respective vertices. Note that for an undirected graph, such as in this case, A is defined to be symmetric.

$$a_{i,j} = \phi(v_i, v_j)$$

PageRank

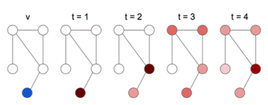
Conceptually, the PageRank algorithm[5] works by assigning each node some initial value. These values are then distributed throughout the network, proportionately to the node's respective edge weights, with an added damping factor (d). This process continues recursively until convergence is reached (see visualization inset below). The result is R, whose values correspond to the PageRank values for each vertex. Once converged, R is static and deterministic for a given graph.

$$R_{t+1} = \frac{(1-d)}{p} + dAR_t$$

This process can be personalized[6] by creating a personalization vector, v, of probabilities for each vertex (example in blue). The PageRank algorithm is started based on that vector. Then, at each iteration, that personalization vector is restarted with some probability (alpha). Similarly, the process continues until convergence.

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

For this purpose, the personalization vector consists of a normalized vector of non-"reason for visit" vertices for a given clinical encounter. The output of the PageRank algorithm is further refined by only taking the "reason for visit" concepts and normalizing them to sum to 1, thus creating a probability distribution for predicted reason for visit only.

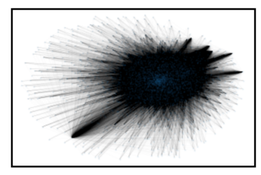


Tuning and Evaluation

Performance was evaluated on the 2015 dataset. This evaluation dataset was preprocessed with the 2014 dataset, but held out of the graph generation and PageRank algorithm. Reasons for visit were predicted for items in the evaluation dataset, and were counted as success if any reason was correctly identified. Alpha was picked using grid search, with performance evaluated on a subset of the evaluation set.

Results

In total, there were 16,846 clinical encounters included in this analysis (13,089 in 2014 and 3757 in 2015). The binarized dataset included 2667 variables, with an average density of 0.5% (range: 0.008% - 89%). This included 657 reason for visit, 1948 medication, 23 diagnosis, and 39 demographic or visit variables.



The graph consisted of 2667 vertices and 383,748 edges. The average unscaled edge weight was 9.8 encounters; the average scaled edge weight was 0.64%. A visualization of the graph (inset) generally shows a few vertices with high degrees of connection, and many that are practically unconnected. This corresponds with quantitative analysis as well: the average degree for nodes 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).

Non-Personalized PageRank Results

Code	Reason	PR
4800.0	Follow-up Visit	0.50%
2510.0	Hypertension	0.38%
4115.0	Medication Check/Refill	0.37%
3100.0	General Medical Exam	0.36%
2205.0	Diabetes Mellitus	0.28%
1440.0	Cough	0.28%
2215.0	Other endocrine disease	0.27%
1905.1	Back Pain	0.25%
1015.0	Tiredness, exhaustion	0.25%
1100.0	Anxiety	0.24%

PageRank was run without any personalization, producing baseline predictions (most probable shown below). A set of prototypical patients were empirically examined, and found modifications to both the predicted probabilities and orderings of predictions. For example, a pregnant woman was predicted to have a "Routine prenatal visit" (probability: 2.1%). The personalized PageRank algorithm correctly identified reason for visits 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. The results aren't spectacular: 31% accuracy leaves significant room for improvement. However, given the large search space, this approach provides a good starting point.

There's plenty of opportunity for improvement in this methodology. Future work could include pruning the edges of the network, or removing highly connected nodes (e.g., seen before and no chronic conditions). This may help to create more distinct sections of the graph, perhaps generating more personalized results. Similarly, removing the most common or generic reasons (e.g., "follow-up visit", "medication check", "general medical exam", "counseling") might allow for more personalized reasons to surface.

More training data, particularly from multiple years, could be used. Unfortunately, there are numerous changes to NAMCS data collection and formats each year. This creates substantial burden in preprocessing this data. Similarly, there are additional data elements (vitals and labs) which could also potentially contribute to this approach. These are typically continuously valued variables, adding additional steps to incorporate them into this graph approach.

Finally, a more detailed evaluation methodology could be used here. Only accuracy was computed. However, further evaluation could be done to see how setting probability thresholds for inclusion (i.e., don't show any predictions below a certain probability) or varying top N reasons might impact performance.

Conclusions

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).

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. Chen J, Yu H. Unsupervised ensemble rankings of terms in electronic health record notes based on their importance to patients. J Biomed Inf. 68 (2017): 121 - 131.
2. Martinez D, Otegi A, Soroa A, Agirre E. Improving search over electronic health records using UMLS-based query expansion through random walks. J Biomed Inf. 51 (2014): 100 - 106.
3. Hristidis V et al. Information Discovery on Electronic Health Records Using Authority Flow Techniques. BMC Medical Informatics & Decision Making. 10 (2010).
4. National Center for Health Statistics. 2014 NAMCS Microdata File Documentation. Centers for Disease Control and Prevention, National Center for Health Statistics: 2019. http://www.cdc.gov/nchs/ahcd.html
5. Page L, Brin S, Motwani S, Winograd T. The PageRank Citation Tanking: Brining Order to the Web. Technical Report 1999-66, Stanford InfoLab, Nov 1999.
6. Haveliwala T et al. An analytical comparison of approaches to personalizing PageRank (Technical Report). Stanford, CA: Stanford University. http:// ilpubs.stanford.edu:8090/596/1/2003-35.pdf

one medical