Identifying Syndromic Fingerprints in Reason Fields in Emergency Department or Telehealth Records using N-grams for Similarity Analysis

El Sayed Mahmoud and Deborah A. Stacey

Computing and Information Science, University of Guelph, Guelph, Ontario, Canada

OBJECTIVE

The objective of this work is to identify syndromic fingerprints in reasons for entering an emergency department (ED) or calling telehealth (TH). It also demonstrates that these fingerprints are valuable for classification.

BACKGROUND

An N-gram is a sub-sequence of *n* items from a given sequence where *n* can be *1, 2,…, n* and the items can be letters or words. N-gram models are widely used in statistical natural language processing [3]. In the syndromic surveillance context, N-grams can be used to cluster or classify natural language data. They can also help in the design of kernels for machine learning algorithms such as support vector machines to learn from text data.

This work calculates the similarity percentages of ED or TH reasons to syndromic fingerprints using N-grams. We define "*reasons similarity*" as the percentage of matched N-grams derived from the reasons field of an ED or TH record with the fingerprint of a syndrome. The *fingerprint* of a syndrome is a list of frequent N-grams related to this syndrome. This fingerprint is constructed by collecting a large sample of classified reasons data for a particular syndrome, calculating all of the N-grams for this set and then selecting the most frequent N-grams to form a profile or fingerprint.

N-gram generation may require extensive processing time especially for large files but this issue has been addressed by using parallel computation.

METHODS

ED and TH reasons were analyzed using N-grams. The letter N-gram method was selected to produce a language independent solution. A parallel program was developed to analyze reasons data and generate N-gram vectors. N was selected to be 5 for our initial research. Our N-gram vector includes the N-grams and their frequencies from a two year (2004-5) collection of classified ED and TH data. N-gram vectors were separately generated for three syndromes: respiratory, rash and gastro. Each resulting N-gram vector was sorted in descending order. The top 25% of the sorted N-grams serve as a fingerprint for the corresponding syndrome. The N-grams of a test set of 100 reasons from each syndrome were classified using these three fingerprints. The similarities between the fingerprints themselves were also studied.

RESULTS

Similarity between the three syndromic fingerprints show that the rash fingerprint is not similar to the others while the respiratory and gastro fingerprints have an 11% similarity (see Figure 1).

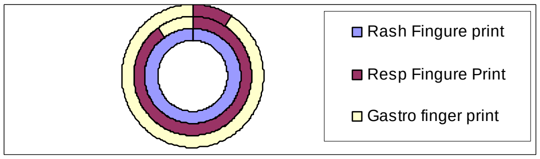

Figure 1: Similarity between fingerprints

When the fingerprints were used to classify the 300 reasons in our test set the average classification accuracy was 87%. Figure 2 illustrates the average similarity of the test data to each of the syndromic fingerprints.

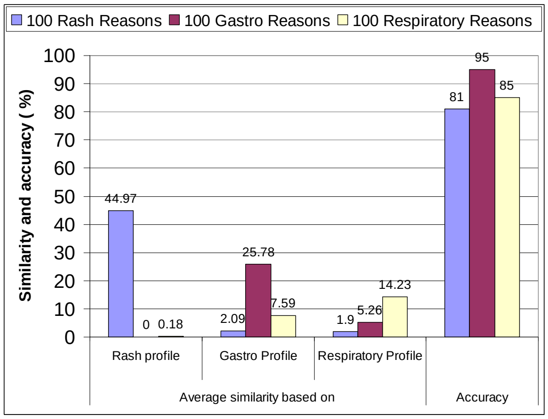

Figure 2: Test Results: Similarity to Fingerprints

CONCLUSIONS

In our experiments, 5-letter grams were enough to generate useful fingerprints for three syndromes. These fingerprints are dissimilar and thus valuable for classification purposes. This type of analysis can be extended to the study of the effect of n-gram length and the possibility of adapting the similarity measure to produce fuzzy or probabilistic classifications.

REFERENCES

[1] W. Cavnar & J. Trenkle (1994) "N-Gram-based Text categorisation", Proceedings of Symposium on Document Analysis and Information Retrieval, Las Vegas, NV.
[2] A Similarity-Based Agent for Internet Searching, Tony G. Rose & Peter J. Wyard (citeseer.ist.psu.edu/rose97similaritybased.html)
[3] http://en.wikipedia.org/wiki/N-gram

Further Information:{emahmoud, dastacey}@uoguelph.ca