AUTOMATED CLASSIFICATION OF ENCOUNTER NOTES IN A COMPUTER BASED MEDICAL RECORD

D.B. Aronow, S. Soderland, J.M. Ponte, Feng F., W.B. Croft, W.G. Lehnert

Center for Intelligent Information Retrieval, Lederle Graduate Research Center, University of Massachusetts, Amherst MA 01003, USA

ABSTRACT

Harvard Community Health Plan is exploring emerging information technologies for means to use the text portion of its 25 year old computerized medical record system. The Center for Intelligent Information Retrieval is developing systems to answer the question: to what extent can automated information systems replace manual chart review of encounter notes? INQUERY, a probabilistic inference net information retrieval system, and FIGLEAF, an inductive decision tree text classifier are applied to the problem of classifying electronic encounter notes to identify acute exacerbations in pediatric asthmatics. Both systems achieve average precisions of greater than 80%, with a new enhancement to INQUERY's relevance feedback the top performer. Refinement of the systems and plans for their integration are discussed.

1. INTRODUCTION

Harvard Community Health Plan (HCHP) is a large Boston-based health maintenance organization, which has benefited from computerized medical records since its founding 25 years ago. The Automated Medical Record System (AMRS) contains essentially the complete medical record more than 300,000 current members [1].

The data within AMRS is a strategic resource for both quality improvement and utilization management. The coded portions are used extensively. However, the text portion of the AMRS resource is inaccessible and virtually ignored, with the exception of resource-intensive manual chart review for quality measurement and research. HCHP is exploring emerging information technologies for means to unlock this data and build into HCHP production systems the capacity to access, extract, manipulate and abstract clinical text data.

The Center for Intelligent Information Retrieval (CIIR) is a National Science Foundation supported State-Industry-University consortium located in the Computer Science Department of the University of Massachusetts in Amherst. The Center was established to focus research in text-based information systems and to facilitate transfer to industry of technologies developed in the department. Two of the systems being developed in this research are an advanced information retrieval system, named INQUERY, that is based on a probabilistic inference net model and a text classification system, named FIGLEAF, that is based on inductive decision tree generation.

1.1 HCHP'S INFORMATION NEED

HCHP joined the CIIR as one means of developing state-of-the-art tools for exploiting clinical text data resources. Identification of acute exacerbations of asthma in children was selected as the pilot project. Asthma is the most common serious childhood illness and the greatest consumer of inpatient resources in children. Several quality improvement (QI) projects have been undertaken at HCHP concerning asthma care, one of which concerns the frequency and circumstances surrounding exacerbations. This QI study has a manual chart review component, part of which requires trained coders to identify AMRS encounter notes in which exacerbations are documented.

The question HCHP posed to CIIR was: to what extent can automated information retrieval replace manual chart review of medical record encounter notes in support of quality measurement. Specifically, can how well can CIIR research systems identify those 5% of encounters in the medical records of pediatric asthmatics which concern acute exacerbations. This is a classification or filtering problem, in that the user is faced with a large collection of documents (medical record encounter notes) of which she wants to consider only those meeting a specified set of characteristics (asthma in acute exacerbation). Two of the research groups in CIIR, the Information Retrieval Laboratory and the Natural Language Processing Group have independently undertaken to solve this problem.

1.2 INQUERY - AN INFERENCE NETWORK INFORMATION RETRIEVAL SYSTEM

The INQUERY system is a text based information retrieval (IR) system. Retrieval is performed in a probabilistic fashion via Bayesian inference networks which combine available evidence, namely the initial query and documents judged relevant by the user, into belief values for each document. The documents are displayed in rank order so that the documents most likely to be relevant are presented to the user first.

A typical application for INQUERY is document retrieval in very large (multiple Gigabyte) text databases. A user presents INQUERY with a query either in INQUERY query language or in natural language and then has the option after viewing the retrieved documents of marking some of them as relevant. INQUERY can then use these relevance judgments to expand the initial query and improve retrieval performance [2].

The current task is a novel one for INQUERY due to the small collection size and the large number of relevance judgments. In addition, the notion of classification as opposed to retrieval is a new task for INQUERY, however the underlying model of probabilistic inference nets is robust enough to provide good performance.

1.3 FIGLEAF - AN INDUCTIVE TEXT CLASSIFICATION SYSTEM

The FIGLEAF (FIne Grained Lexical Analysis Facility) text classification system, currently under development, bases its classification on decision trees derived from examples in a set of training documents [3]. Each document to be classified is encoded as a list of feature-value pairs and is passed to an ID3 decision tree, which returns a probability that the document is a positive instance. Rather than considering every word in a document as a feature, FIGLEAF uses a dictionary of words highlighted by a domain expert, who marks the words or phrases in each training text that were used in determining its classification. FIGLEAF ignores all words not in its master dictionary and any section of the document never used during text marking.

In addition to terms and bigrams (pairs of adjacent terms) from training documents, lists of highly predictive, conceptually related terms can be turned into "meta-features" whose values are the counts of how many of those terms occur in a document. Meta-features can be supplied by domain experts or automatically created by tabulating occurrences of terms and bigrams in positive and negative training instances.

Decision trees are automatically induced from training instances using the ID3 decision tree algorithm. Each node in the decision tree is a test for the value of a feature. ID3 selects the test that most effectively separates positive and negative training instances, using an information gain metric [4]. Features are recursively selected to partition the training instances. After the tree has been built, pruning off branches near the leaves generally improves performance. Pruning thresholds must be empirically determined for each tree.

The implementation of ID3 used by FIGLEAF associates each node of the decision tree with a confidence level from 0.0 to 1.0, based on the proportion of training instances at that node which are positive. These confidence levels are then used to rank-order documents.

2. METHODOLOGY

2.1 DATA

Using coded data fields in AMRS, a pediatric asthmatic population was identified comprised of all current members under age 18 with two or more visits for asthma. From this group, a training-data group and a testing-data group were defined, made up of 75 and 25 randomly selected patients respectively. Each patient’s electronic medical record was then filtered to exclude all encounter notes containing no occurrence of a code for asthma or asthma-like conditions (Acute Bronchitis, Bronchiolitis and Bronchospasm), or where no definitive assessment or diagnosis was made (Diagnosis Deferred). Each encounter note was classified by a clinician as either relevant if an acute asthmatic exacerbation, or irrelevant if any other encounter.

By this process, a training collection, PA_TRAIN, and testing collection, PA_TEST, were created which represent the most refined fully automated selection of encounter notes from which manual reviews can identify acute exacerbations. PA_TRAIN contains 988 encounter documents, of which 293 (29.7%) are acute exacerbations, while

PA_TEST contains 260 encounters, 60 (23.1%) of which are acute exacerbations. Before the automated code-based filter, 4.4% of both training and testing collections were acute exacerbations

2.2 INQUERY

It has been shown that a variety of user defined conceptual query paths can be used on pediatric asthma data with varying degrees of success [5]. For the current experiment, the best previously identified conceptual query path, that of asthma medications, route and effects, was used. The initial query was run against PA_TEST to establish a baseline result. The query was also run on PA_TRAIN using relevance feedback with all the relevant training collection documents to create an expanded query.

It is a well established result of the machine learning community that negative information e.g. irrelevant documents, is helpful for classification. The relevance feedback algorithm used by INQUERY uses only positive information for real-time retrieval tasks. Since the entire PA_TRAIN has been classified, and there was ample negative evidence available, we have devised a method to incorporate that evidence in the current experiment.

In order to include that negative evidence, a query consisting of just the word "asthma" was run on PA_TRAIN and relevance feedback was done using the irrelevant documents in the collection, instead of the relevant documents. The resulting “negative” expanded query terms were combined using the INQUERY probabilistic NOT operator with the usual, “positive” expanded query terms obtained from relevance feedback of the initial medication query run on PA_TRAIN. Finally, the fully expanded query, with both “positive” and “negative” evidence included, was run on PA_TEST.

2.3 FIGLEAF

The basic FIGLEAF instance encoding was comprised of features for individual terms obtained from clinician highlighted words during text-marking, and bigram features for pairs of adjacent terms such as CHEST-CLEAR or EMERGENCY-ROOM,. Additional meta-features were created from terms and phrases supplied by five of the conceptual query paths (asthma medications, routes, effects, as well as symptoms and hospitalization) previously prepared for the asthma domain [5]. Finally, statistical analysis was used to define lists of terms most highly correlated with relevant and with irrelevant documents, forming the basis of further meta-features.

After constructing one ID3 decision tree from PA_TRAIN with meta-features and another tree without meta-features, optimal thresholds for pruning each tree were found by ten-fold cross-validation. Both trees were then applied to PA_TEST and their independent confidence levels were combined to produce a final ranking of test documents as positive instances of asthma exacerbation.

3. RESULTS

As shown in Table 1 and Figure 1, INQUERY query expansion using both positive and negative evidence produced a marked improvement in precision compared to the best conceptual path initial query. To confirm the power of this query expansion approach, recipes and popular song lyrics were also used as initial queries, with no significant difference in average precision. FIGLEAF achieved performance better than the initial INQUERY query, but did not equal the INQUERY expanded query.

4. DISCUSSION

4.1 INQUERY

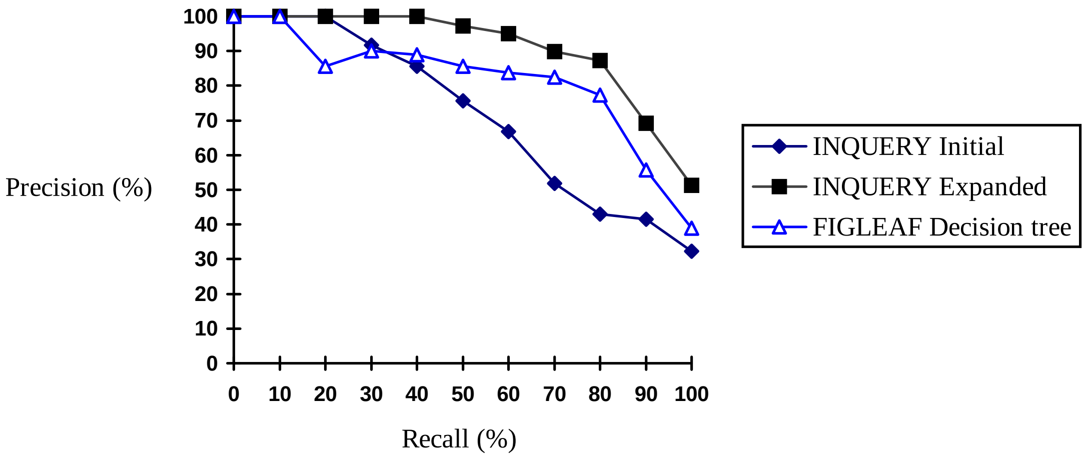
The effectiveness of including both positive and negative information in relevance feedback is encouraging in this experiment. The specific content of the initial query had minimal effect on performance, suggesting that relevance feedback in the probabilistic inference net model is a good choice for classification of medical record encounter notes. Further experiments need to be done on other document and query sets to insure that this performance can be duplicated. In addition, experiments using positive and negative information from a smaller set of relevance judgments need to done to test the effectiveness of the technique in situations where a large quantity of training data is not available

Table 1: Recall and precision values for INQUERY and FIGLEAF experiments

	INQUERY	query	FIGLEAF

	Initial	Expanded	Decision tree
Recall	Precision	Precision	Precision
0	100.0	100.0	100.0
10	100.0	100.0	100.0
20	100.0	100.0	85.7
30	91.7	100.0	90.0
40	85.7	100.0	88.9
50	75.6	97.2	85.7
60	66.7	95.0	83.7
70	51.9	89.8	82.4
80	42.9	87.3	77.4
90	41.5	69.2	55.7
100	32.3	51.3	39.0
Average	71.7	90.0	80.8

Fig.1 Recall - Precision Curves for INQUERY and FIGLEAF



4.2 FIGLEAF

FIGLEAF's performance on this data set was somewhat worse than we had seen in earlier experiments based on a slightly different HCHP training corpus (approximately 50% relevant, rather than half that value, as are the current data sets). A weakness in the ID3 decision trees used in FigLeaf appear to be their sensitivity to noise. Discriminations made near the leaf nodes were often based on incidental characteristics of the training data and classification could be improved significantly by pruning the tree, cutting off the smaller branches. Unfortunately the optimal pruning found empirically for the training data is not necessarily optimal for performance on the test set.

Combining evidence from two qualitatively different decision trees went a long way towards reducing noise sensitivity and dependence on optimal pruning. When one tree incorrectly gave high (or low) confidence to an instance due to noise, the other tree would generally damp this effect with a mid-range confidence value.

Aspects of FIGLEAF that require human expertise can be replaced with statistical analysis of the training set with little degradation in performance. Statistically derived meta-features are nearly as effective as those from conceptual query paths. The dictionary itself can be built from terms correlated with either relevant or irrelevant training documents, avoiding the labor of text-marking.

5. CONCLUSIONS

HCHP seeks to reduce the time and cost of chart review in health care quality improvement projects through automated classification of encounter notes. For HCHP, the goal is to correctly classify as many documents as possible. This differs from a typical retrieval task which provides good documents at the top of a belief list by achieving high precision at the low recall end of the recall/precision curve. Thousands of documents might satisfy the information need of the user, but the user only has time to look at the best few. Encounter note classification, on the other hand, requires high recall, as the user wants to miss as few relevant notes as possible, even at the potential cost of receiving a large number of irrelevants.

Encounter note classification, either using an INQUERY or FIGLEAF based system, appears quite promising from these preliminary experiments. INQUERY performed better than FIGLEAF in the experiments reported here and requires far less time to train. Enhancements to the relevance feedback features of INQUERY continue to sharpen its performance. FIGLEAF is still in an early development stage and methods of automating the training process for FIGLEAF are being studied. More definitive conclusions about the relative performance of these techniques for classification of medical text, however, require testing on a number of different clinical topics.

We plan to explore integration of the systems into a two-stage filtering process. In one approach, INQUERY might provide, via automated query expansion, candidate dictionary terms to FIGLEAF. In a second approach, one system could be tuned to serve as an initial, high recall/low precision filter which passes selected documents to the second system which would tuned for low recall/high precision.

Ultimately, the output from these systems must better fit the information needs of the user. Efficiencies in chart review can most immediately be gained by reducing the number of records which must be manually reviewed. To do this, the belief rankings produced by the systems will need to be converted into probability ranges and the output presented to the user as a three-bin sort: Positive, Uncertain and Negative. Confidence levels should be tunable by the user, so that for any particular quality measurement project, the user can define how broad a range of notes (Uncertains) to manually review.

6. ACKNOWLEDGMENTS

This material is based on work supported in part by NRaD Contract Number N66001-94-D-6054. This material is also based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623 and by the Harvard Community Health Plan Foundation. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

7. REFERENCES

1. Schoenbaum SC, Barnett GO. Automated ambulatory medical records systems: An orphan technology. *International Journal of Technology Assessment in Health Care*. 8:598-609, 1992.
2. Broglio J, Callan JP, Croft WB. INQUERY system overview. In: *Proceedings of the TIPSTER Text Program (Phase I*). San Francisco, CA: Morgan Kaufmann, 1994, pp 47-67.
3. Lehnert W, Soderland S, Aronow D, Feng F, Shmueli A. Inductive text classification for medical applications. *Journal of Experimental and Theoretical Artificial Intelligence*, (In Press), 1994.
4. Quinlan, JR. Induction of Decision Trees. *Machine Learning* 1:81-106, 1986.
5. Aronow DB, Callan JP, Croft WB, Ponte JM. Document filtering for quality measurement using an inference network retrieval system. *Eighteenth Annual Symposium on Computer Applications in Medical Care*. (In Review).