IDENTIFYING HIGH QUALITY MEDLINE ARTICLES AND WEB SITES USING MACHINE LEARNING

By

YINDALON APHINYANAPHONGS

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Informatics

December 2007

Nashville, Tennessee

Approved:

Professor Constantin Aliferis

Professor Ioannis Tsamardinos

Professor Douglas Hardin

Professor Steven Brown

Professor Dan Masys

DEDICATION

My parents. My family.

ACKNOWLEDGEMENTS

You could say that I had the best twenties a person could probably have. I got to wake up every morning excited and motivated to discover new knowledge and solve real problems, and at the same time, develop as a person and a scientist. I'm ready for the world and ready to be an adult.

The tops of my list to thank are my parents and family. Without the consistent pressure of my mom asking how the dissertation was going, when I was going to be done, and what I was going to do next, I'm not quite sure I'd be writing this today. My parents are the source of who I am, and I thank them for making all the sacrifices so I could pursue my dream.

Next is my advisor Constantin Aliferis. I could not ask for a better mentor and friend, and a clear sign of my advocacy for Dr. Aliferis as an advisor is that I would have him all over again, and I would not hesitate to recommend him to any student.

The measure of any mentor is does the mentor prepare you to make contributions to the field and does the mentor teach you how to be a good scientist. My answer is an unequivocal yes to both questions. I feel prepared to make contributions on my own to the field and I have learned how to do "good" science.

I'd like to thank all my committee members. Dr. Tsamardinos, Dr. Brown, Dr. Hardin, and Dr. Masys for their feedback and involvement. I'd also like to thank the MSTP department, the NLM training fellowship, and the Department of Biomedical Informatics for supporting me.

There are a few other people I have to thank. My little brother Joe for being a hard ass and keeping me focused. My buddy Michael and JP for giving me refuge for a couple of days in the final stretch of dissertation writing so I could finish. Sutin for discussing software engineering with me and taking me away from the world of medicine every so often, and keeping me inspired. Mimmie, Micaela, Todd, and Jeff for listening to me rant and rave during my toughest emotional times so I could get it out of my system for the day to focus on research.Geoff for being supportive and a great friend, and keeping me in the loop on the outside world. Finally, all my peeps that left me to pursue their dreams. Jeff, Troy, Darcie, Allison, and Mary Hunt. If it wasn't for their continued friendship, I would not have been able to make it to where I am today.

TABLE OF CONTENTS

DEDICATION .... **iii**

ACKNOWLEDGEMENTS .... **iv**

LIST OF TABLES .... **ix**

LIST OF FIGURES .... **xi**

I. INTRODUCTION **1**

OVERVIEW AND DISSERTATION STRUCTURE 4

Models and Evaluation of Retrieval Performance 5
Evaluation of Generalization 6
EBMSearch: Proof of Concept Search Engine 7
Extensions to the World Wide Web 7
Conventions 8
SUMMARY 8

II. BACKGROUND/ PRIOR WORK **9**

TEXT CATEGORIZATION MODELS FOR HIGH QUALITY RETRIEVAL IN INTERNAL MEDICINE 9

Abstract 9
Introduction 11
Background 12
Methods 15
Results 27
Discussion 35
Summary 42
References 43
LEARNING BOOLEAN QUERIES FOR ARTICLE QUALITY FILTERING 46

Abstract 46
Introduction 46
Methods 49
Results 55
Discussion 59
Conclusions 62
References 63

III. MODELS AND EVALUATION OF RETRIEVAL PERFORMANCE **65**

A COMPARISON OF CITATION METRICS TO MACHINE LEARNING FILTERS FOR THE IDENTIFICATION OF HIGH QUALITY MEDLINE DOCUMENTS 65

Abstract 65
Introduction & Background 67
Hypothesis & Experiments 69
Methods 70
Results 81
Study Limitations 90
Discussion and Conclusions 95
References 98
A COMPARISON OF WEB HYPERLINKS TO MACHINE LEARNING FILTERS FOR THE IDENTIFICATION OF HIGH QUALITY MEDLINE DOCUMENTS 105
Abstract 105
Introduction 106
Background 108
Methods 110
Results 118
Study Limitations 120
Discussion and Conclusions 121
References 125

IV. EVALUATION OF GENERALIZATION OF MACHINE LEARNING MODELS 128

PROSPECTIVE VALIDATION OF TEXT CATEGORIZATION FILTERS FOR IDENTIFYING HIGH-QUALITY, CONTENT-SPECIFIC ARTICLES IN MEDLINE. 128
Abstract 129
Introduction 129
Methods 132
Results 138
Discussion 142
References 143
EXTENDING TEXT CATEGORIZATION FILTERS IN MEDICINE 146
Abstract 146
Introduction 147
Hypothesis 151
Methods 152
Discussion 176
Limitations 179
Conclusions 180
References 180

V. EBMSEARCH: PROOF OF CONCEPT SEARCH ENGINE 184

EBMSEARCH 184

VI. EXTENSIONS TO THE WORLD WIDE WEB 191

TEXT CATEGORIZATION MODELS FOR IDENTIFYING UNPROVEN CANCER TREATMENTS ON THE WEB 191
Abstract 191
Introduction 192
Hypothesis 195

Materials and Methods 196
Results 201
Discussion..202
Limitations..205
Conclusions 205
References 206

VII. CONCLUSIONS 209

OPEN PROBLEMS 210

LIST OF TABLES

Table II-1 - Clinical Query Filters described in the "filter table" used in the clinical queries link in PubMed (3). These Boolean filters were run on the gold standard corpus and sensitivity, specificity, and precision were measured............................ 14

Table II-2 - Best learning method compared to clinical query filters fixed at optimal sensitivity and specificity. The first number is the average across 5 folds. The numbers in parenthesis report the minimum and maximum value across the 5 folds. Cells in bold denote the performance for the filter optimized for sensitivity or specificity respectively... 26

Table II-3 - Area under the Receiver Operating Curve (AUC) Performance of each machine learning method in each category... 29

Table II-4 – McNemar’s test p-values averaged over 5 folds with significance tests. The permutation significance is produced by random permutation tests as described in Section IIIF. 34

Table II-5 – Feature Selection Performance 56

Table II-6 – Decision Tree Performance on Test Set 57

Table II-7 – Decision Trees Compared to CQF filters 58

Table II-8 – Etiology Decision Tree Comparison.. 61

Table III-1 – Comparison of gold-standard-specific, content-based machine learning filters with citation metrics and models built for ACPJ criteria in the SSOAB quality classification task 81

Table III-2 – Features selected by the HITON_PC algorithm from the entire SSOAB corpus (i.e. using all documents). Some words are stemmed [15]. 84

Table III-3 – Comparison of gold-standard-specific, content-based machine learning filters with hybrid content + citation metric models 87

Table III-4 – Results of Support Vector Regression prediction of citation counts from content compared to the true citation counts in all corpora. 89

Table III-5 – Comparison of ACPJ_etiology-specific content-based machine learning filters with citation metrics in the ACPJ-Treatment quality classification task........ 91

Table III-6 - Comparison of ACPJ_treatment-specific, content based machine learning filters with citation metrics in the ACPJ-Treatment quality classification task........ 92

Table III-7 - Positives and negatives distribution in each category. The sizes were limited due to Google and Yahoo API restrictions. 117

Table III-8 - Area under the ROC curve for each ranking method. *- area under the curve below 0.5 indicates that the method ranks articles in reverse order. Thus reversing the ranking would rank the positives higher than the negatives. 118

Table III-9 - Average area under the ROC curves for each method............................. 119

Table IV-1 - Top row is cross-validation estimated area under the curve for optimal 1998-1999 and 1998-2000 models. Bottom row is area under the curve for the optimal models applied to 2005 corpora (*no cross-validation applied*). Treat – treatment, Diag – diagnosis, Prog-prognosis, Etio – etiology. ± - is the range of AUC estimates across the 5 folds... 139

Table IV-2 – Optimized Support Vector Machine (SVM) compared to Clinical Query Filters fixed at optimal sensitivity and specificity. All values are calculated using the entire 2005 corpora. 141

Table IV-3 – Labeling Criteria 153

Table IV-4 – 79 out of 161 Randomly Selected Journals Reviewed for Hedges Corpora. A more comprehensive list is available from the authors.................................... 158

Table IV-5 – AUC for specific categories.. 169

Table IV-6 – Sensitivity/ Specificity Optimized CQF Comparison 171

Table VI-1 – Area Under Curve for Each Discrimination Method 202

Table VI-2 - Web page excerpts where previous tools fail to detect unproven claims. A page that makes unproven claims is identified as such if it has a small support vector machine rank, a large quackometer score, and a large Google rank, respectively. SVM rank is calculated over 10 fold cross validation test set composed of 9 positives and 9 negatives. Google rank is out of the top 30 results returned. Quackometer score provides ranks from 0 to 10. “S” denotes success of the corresponding filter, while “F” failure.. 204

LIST OF FIGURES

Figure II-1 - Receiver Operating Curves for Each Category 30

Figure II-2 – 11 Point Precision Recall curves compared to optimized sensitivity and specificity clinical query filters. 31

Figure II-3 – Title+Abstract (TA) vs. Title+Abstract+MeSH+Publication Types (TAM) Performance Comparisons. 40

Figure II-4 – Experimental Design Methodology.. 55

Figure II-5 – Decision tree produced by HITON/ DT 56

Figure III-1 – Average HITS curves used on SSOAB corpus. The GSS SSOAB model returns the most documents in the first 150 articles. Citation count and the NSS ACPJ Txmt Model applied to the SSOAB corpus return fewer documents in the top 150 returns. The SSOAB corpus was composed of 431 positives and 7,379 negatives. Each curve was an average across all folds. 103

Figure III-2 – Average precision-recall curves used on SSOAB corpus. The GSS SSOAB model returns the best performing precision-recall curve. Citation count and the NSS ACPJ Txmt Model have curves below, thus performing lower than, the GSS SSOAB model. The SSOAB corpus was composed of 431 positives and 7,379 negatives. Each curve was an average across all folds. 104

Figure IV-1 - Format Category Receiver Operating Curves 165

Figure IV-2 - Rigor Category Receiver Operating Curves... 166

Figure IV-3 - Purpose Category Receiver Operating Curves (Cost, CPG, Diagnosis, Etiology) 167

Figure IV-4 - Purpose Category Receiver Operating Curves (Economics, Prognosis, Qualitiative, Treatment)... 168

Figure IV-5 – Start page for EBMSearch.org... 174

Figure IV-6 – Results Page for EBMSearch.org 175

Figure V-1 – Index page for EBMSearch system... 185

Figure V-2 – Results page of EBMSearch system... 187

Figure V-3 – Proof of concept interface. The left side will display the MeSH tree if one of the terms is a MeSH term. Suggestions are made if the search query matches part of the MeSH term vocabulary. For example, “diabetes mellitus” and “diabetes insipidus” are suggested for a search of “diabetes.” 189

Figure VI-1 - Receiver operating curves for each method. 202

CHAPTER I

I. INTRODUCTION

We are inundated by information. Cellular phones, personal digital assistants, radios, TV, advertisements on billboards, email, the internet, etc, have all contributed to keeping us connected and keeping inputs of information flowing toward us. Having too much information can be bad. With too much information, we may be exposed to contradictions in views and low signal to noise ratios that make it increasingly difficult to find high quality information to make a relevant decision. Enrico Coiera [1] postulates an impending “information famine” based on Malthus’ law, where, rather than human population needs outgrowing their food sources, the information glut outgrows humans limited ability to find and assimilate high quality information. Malthus’ predictions of widespread famine did not come to pass. Malthus did not foresee the vast advances in agricultural technology that would feed the world’s population. In a similar vein, I propose that advanced search technology may provide a solution to the current information glut. If we can increase the *accuracy* of information search technology at a greater rate than information grows, we may avoid the forthcoming “information famine.”

The “information famine” becomes dire as the health professions increasingly embrace the premise of evidence based care. Virtually every publication and talk regarding the biomedical literature mentions its current volume and exponential growth, and the growing challenge for health professionals and medical librarians in identifying high quality articles applied to evidence based care effectively and efficiently.

Health professionals justify medical decisions through the medical literature, and finding research articles to support a medical decision has become a prelude to better care and health outcomes. Unfortunately, the number of research articles continues to grow so fast that finding the research article needed is increasingly difficult.

The “information glut” on the web is even worse. The lack of quality standards and ease of publishing allow a wide range of quality information [2]. Studies have shown that health consumers have limited ability to evaluate information on medical web sites [3, 4]. Fortunately, to this date, there have been limited examples of adverse outcomes either by health consumers or professionals due to wrong information found on the web [5, 6]. However, as the use of web resources grows the potential for such outcomes increases.

Machine learning, specifically text categorization, provides a solution to identifying documents or websites that match quality standards. With a moderately small set of manually classified documents or websites, a text classification algorithm learns an applicable statistical model. I present a novel use of text categorization in medicine to solve practical problems in identifying quality documents or web pages.

The motivation of the following work is to extend original work using text categorization for identifying high quality articles in internal medicine. The goal is to show that the machine learning filter models built for a specific task compare favorably to other methods to identify quality articles, and the machine learning filter models built for a specific task generalize to time periods outside the training time period. The machine learning filter models can also be used in specific content categories and areas outside of internal medicine. I also present an implementation of these models for identifying high quality articles in all of MEDLINE.

In addition, I apply similar text categorization methods to identify low quality web sites. The framework was used to solve practical problems in identifying articles and web pages in medicine.

Overview and Dissertation Structure

At the core of this dissertation is the application of machine learning pattern recognition techniques, specifically text categorization, for identifying information in the medical literature and the web. The dissertation follows a logical progression of experiments.

In previous work I provided context and motivation for the proposed work. I showed that using powerful text categorization techniques and a suitably constructed, high quality and content labeled article collection for training, one can automatically construct quality filters to identify articles in the content areas of treatment, prognosis, diagnosis, and etiology in internal medicine that perform with better sensitivity, specificity, and precision than current methods. I also showed that it is possible to automatically construct Boolean queries from a corpus using machine learning techniques such that the Boolean queries have as good classification performance as the SVM models, and the resulting Boolean queries are human-readable, manageable, and simple for use in current search engines.

The dissertation is composed of 4 parts – namely *MODELS AND EVALUATION OF INFORMATION RETRIEVAL PERFORMANCE*, *EVALUATION OF GENERALIZATION*, *EBMSEARCH PROOF OF CONCEPT SEARCH ENGINE SYSTEM*, and *EXTENSIONS TO THE WORLD WIDE WEB*. "Models and evaluation of information retrieval performance" compares the machine learning models to other citation and web based measures of identifying quality articles. "Evaluation of generalization" identifies time periods, content categories, and areas outside of internal medicine where the machine learning models are applicable. "EBMSearch proof of concept search engine system" implements the models in a proof of concept system. Finally "extensions to the World Wide Web" take the general machine learning framework and apply it to identifying quality web pages on the internet.

Models and Evaluation of Retrieval Performance

In this section, I studied the machine learning models and compared to other methods to identify high quality articles in the literature. I presented comparisons of specific machine learning filter models built for a specific gold standard to bibliometric citation count, impact factor, and non-specific machine learning models for identifying high quality articles. Furthermore, with the growth of medical content and secondary sources of medical information on the web and based on the observation that higher quality articles should be cited more often and on better websites than lower quality articles, it may be possible to use metrics such as Google PageRank or Yahoo WebRanks to rank the medical literature. I explored this possibility and compared the discriminatory power of web measures to specific machine learning filter models built for the specific gold standard.

The machine learning models built for a specific gold standard outperformed bibliometric citation count, impact factor, non-specific machine learning models, Google PageRank, and Yahoo WebRanks in identifying articles from a constructed gold standard. Specific machine learning filter models were superior to other methods in identifying articles in the literature.

Evaluation of Generalization

In this section, I explored the generalization of the machine learning models for identifying high quality articles in another time period outside the time period for articles used to train the models, to areas outside of internal medicine including pediatrics, oncology, and surgery, and to format, purpose, and rigor content categories.

In the first set of experiments, I built models using a 1998-2000 gold standard and evaluated the models' ability to identify high quality articles in a labeled 2005 gold standard in the content categories of treatment, diagnosis, prognosis, and etiology. I found that the models built using previous years identified articles in the 2005 dataset with area under the receiver operating curve upwards of 0.94. The selected gold standard is a stable, reliable gold standard and the machine learning methodology provides robust models and model performance estimates. Machine learning filter models built with the 1998-2000 corpus can be applied to identify high quality articles in another time period.

In the next set of experiments, I expanded the gold standard to include labeled articles in other areas of medicine such as pediatrics and surgery, format categories such as original, review, case reports, and general/ miscellaneous articles, purpose categories including etiology, prognosis, diagnosis, treatment, costs, economics, clinical prediction guide, and qualitative content, and rigor categories for clinical prediction guide and economics content. The models using this labeled dataset had estimated performances upwards of 0.94 area under the receiver operating curve in identifying articles in other areas of medicine in the format, purpose, and rigor categories. Machine learning models generalize effectively to identify articles in several content categories and other areas of medicine.

EBMSearch: Proof of Concept Search Engine

In this section, I implemented a proof of concept system called EBMSearch that applies the machine learning filter models to a subset of MEDLINE articles. Models were built for 4 categories and applied to a subset of MEDLINE articles published from 2000 to 2006. I developed a simple interface that accepted a Pubmed query, content category, and time period and returned a list of articles ranked by scores output from the models.

Extensions to the World Wide Web

In this section, I extended the general machine learning framework to identify web pages that make false cancer treatment claims. Patients with conditions that are not currently fully treatable are susceptible to unproven and dangerous promises about miracle treatments. In extreme cases, fatal adverse outcomes have been documented. To help protect patients, who may be desperately ill and thus prone to exploitation, I explored the use of machine learning techniques to identify web pages that make unproven claims. The resulting models identify web pages that make unproven claims in a fully automatic manner, and substantially better than previous web tools and state of the art search engine technology.

Conventions

In this dissertation, “I” and “we” are used where appropriate. Published chapters are collaborative in nature, and “we” is used. In other unpublished chapters, the inserted papers are manuscripts intended for journal submission, and “we” is used. In core dissertation content, “I” is used.

Each chapter is a published paper or manuscript. Each chapter has its own references that may refer to citations that are chapters in this dissertation. The text and references of published chapters is retained to maintain the original copyright of the published works.

Summary

This dissertation focuses on increasing the accuracy of information search technology by applying pattern recognition techniques to identify quality and content both in the medical literature and the web. I evaluated the models built by the pattern recognition techniques against citation metrics (bibliometric citation count and impact factor) and web link metrics (Google PageRank and Yahoo WebRanks). Furthermore, I generalized the machine learning models to time periods outside the time period used to build the models, other areas of medicine including oncology, pediatrics, and surgery, and other format, purpose, and rigor content areas. Next, I presented a proof of concept system that implements the models. Finally, the pattern recognition framework was extended for use on the web to identify web pages that make false cancer treatment claims.

The following two chapters summarize initial work that served as the background for this dissertation.

CHAPTER II

II. BACKGROUND/ PRIOR WORK

This chapter reviews text categorization and it use in medicine. In the first set of experiments, I present experiments as relevant background that describes and evaluates the use of text categorization in medicine. In a second set of related experiments, I present a method to convert the text categorization models to related, relevant Boolean queries.

Text Categorization Models for High Quality Retrieval in Internal Medicine

Aphinyanaphongs Y, Statnikov A, Tsamardinos I, Hardin D, Aliferis, C. "Text Categorization Models for High Quality Article Retrieval in Internal Medicine." J American Medical Informatics Association. 2005; 12 (2): 207-216.

Abstract

OBJECTIVE: Finding the best scientific evidence that applies to a patient problem is becoming exceedingly difficult due to the exponential growth of medical publications. The objective of this study was to apply machine learning techniques to automatically identify high quality, content-specific articles for one time period in internal medicine and compare their performance to the Boolean-based PubMed clinical query filters of Haynes, et. al.

DESIGN: The selection criteria of the ACP Journal Club for articles in internal medicine were the basis for identifying high quality articles in the areas of etiology, prognosis, diagnosis, and treatment. Naïve Bayes, a specialized AdaBoost algorithm, and linear and polynomial support vector machines were applied to identify these articles.

MEASUREMENTS: The machine learning models were compared in each category to each other and to the clinical query filters using area under the receiver operating characteristic curves, 11-point average recall-precision, and a sensitivity/ specificity match method.

RESULTS: In most categories, the data-induced models have better or comparable sensitivity, specificity, and precision than the clinical query filters. The polynomial support vector machine models perform the best among all learning methods in ranking the articles as evaluated by area under the receiver operating curve and 11-point average recall-precision.

CONCLUSIONS: This research shows that, using machine learning methods, it is possible to automatically build models for retrieving high quality, content-specific articles, using inclusion or citation by the ACP Journal Club as a gold standard, in a given time period in internal medicine that perform better than currently-used PubMed clinical query filters.

INDEX TERMS: Information Retrieval, PubMed, Artificial Intelligence, Machine Learning

Introduction

Evidence Based-medicine (EBM) is an important development in clinical practice and scholarly research. The aim of EBM is to provide better care with better outcomes by basing clinical decisions on solid scientific evidence. EBM involves three distinct steps: (a) identification of evidence from the scientific literature that pertains to a clinical question, (b) evaluation of this evidence, and (c) application of the evidence to the clinical problem [1].

In practice, the application and adoption of EBM to real life clinical questions is challenging. Insufficient time for searching, inadequate skills to appraise the literature, and limited access to relevant evidence are among the most cited obstacles. Coupled with the scientific literature's exponential growth, applying EBM in daily practice proves a challenging and daunting task [2]. This paper addresses the barriers to EBM by improving physician access to the best scientific evidence, (i.e. the first step of EBM).

We hypothesize that by using powerful text categorization techniques and a suitably constructed, high quality and content labelled article collection for training, we can automatically construct quality filters to identify articles in the content areas of treatment, prognosis, diagnosis, and etiology in internal medicine that perform with better sensitivity, specificity, and precision than current Boolean methods. We note that throughout this paper, references are made to both full-text articles and MEDLINE records. We clarify that (a) our filters make judgments about articles and (b) these judgments are made using the MEDLINE records (i.e. titles, abstracts, journal, MeSH terms, and publication types) as the latter are provided by PubMed. Hence, when the context is about processing the records we use "MEDLINE records", whereas when we discuss making judgments about the articles we use the term "articles".

The background section describes previous approaches for identifying the best scientific evidence. The methods section describes corpus construction, the representation of an article (i.e. as a MEDLINE record), articles that meet rigorous EBM standards (high quality) and those that do not, and the learning methods applied to differentiate high quality articles from articles that do not meet EBM criteria. In the results and discussion, we compare the machine learning methods to each other using ROC analysis and 11 point precision recall and to current methods with standard sensitivity, specificity, and precision metrics, and a sensitivity/specificity match method. We further discuss advantages, limitations, and extensions of this work. We conclude with a broad overview of the findings of this paper.

Background

Specialized sources for high quality scientific evidence include The Cochrane Collaboration's Library, *Evidence-Based Medicine*, and the *ACP Journal Club* [3-5]. Each group and journal brings together expert reviewers who routinely review the literature and select articles that warrant attention by clinicians. These articles are either cited by the Cochrane Collaboration, or republished with additional commentary as in *Evidence-Based Medicine* and the *ACP Journal Club*.

These manual methods are labor-intensive and the reporting of high quality articles is slow due to the expert review process. In light of these limitations, more recent approaches address finding high quality, content specific articles as a classification problem. The problem is to classify documents as both high-quality and content-specific or not.

In 1994, Haynes and colleagues used the classification approach to find high quality articles (as represented by their MEDLINE record) in internal medicine [6]. Evaluating articles in ten journals from 1986 and 1991, three research assistants defined high quality articles by constructing a gold standard according to content and methodological criteria. The content areas included etiology, prognosis, diagnosis, and treatment, and the methodological criteria were similar to the criteria currently used by the ACP Journal Club [7]. The authors selected terms that would most likely return high quality articles in these content categories based on interviews with expert librarians and clinicians. Valid MeSH terms, publication types, and wildcarded word roots (i.e. random* matching *randomize* and *randomly*) in the title and abstract were collected. Using the above gold standard and the selected terms, they ran an exhaustive search of all disjunctive Boolean set term models of 4 to 5 terms, and evaluated each disjunctive set on an independent document set according to sensitivity, specificity, and precision of returning high quality articles. The optimal Boolean sets (see Table II-1) were shown to have high sensitivity, specificity, and precision and are currently featured in the clinical queries link in PubMed [8]. This method required interviewing to select terms, a gold standard constructed by an ad-hoc review panel of expert clinicians, and reliance on NLM assigned terms. The learning method also relied on a search of term disjunctions that grows exponentially with the number of search terms.

Table II-1 - Clinical Query Filters described in the "filter table" used in the clinical queries link in PubMed (3). These Boolean filters were run on the gold standard corpus and sensitivity, specificity, and precision were measured.

Category	**Optimized for**	**PubMed equivalent**
Therapy	sensitivity	"randomized controlled trial" [PTYP] OR "drug therapy" [SH] OR "therapeutic use" [SH:NOEXP] OR "random*" [WORD]
	specificity	(double [WORD] AND blind* [WORD]) OR placebo [WORD]
Diagnosis	sensitivity	"sensitivity and specificity" [MESH] OR "sensitivity" [WORD] OR "diagnosis" [SH] OR "diagnostic use" [SH] OR "specificity" [WORD]
	specificity	"sensitivity and specificity" [MESH] OR ("predictive" [WORD] AND "value*" [WORD])
Etiology	sensitivity	"cohort studies" [MESH] OR "risk" [MESH] OR ("odds" [WORD] AND "ratio*" [WORD]) OR ("relative" [WORD] AND "risk" [WORD]) OR "case" control*" [WORD] OR case-control studies [MESH]
	specificity	"case-control studies" [MH:NOEXP] OR "cohort studies" [MH:NOEXP]
Prognosis	sensitivity	"incidence" [MESH] OR "mortality" [MESH] OR "follow-up studies" [MESH] OR "mortality" [SH] OR prognos* [WORD] OR predict* [WORD] OR course [WORD]
	specificity	prognosis [MH:NOEXP] OR "survival analysis" [MH:NOEXP]

PTYP – publication type	MESH – MeSH main heading
SH – MeSH subheading	NOEXP – MeSH subtree for the term is not exploded

Other researchers have applied a similar methodology to developing sets of search terms for controlled trials, systematic reviews, and diagnostic articles [9] [10] [11] [12] [13] [14].

The common methodological features of these studies are: (a) that the search term sets are selected through interviews or article inspection by health professionals and/or librarians and (b) search is conducted via Boolean queries involving combinations of MeSH qualifiers, MeSH terms, publication types, and text words. The selection of a gold standard varies with more recent research utilizing reproducible, expert-derived gold standards. In the present research, we follow an expert-derived, publisher-based methodology for gold standard construction while automating term selection from the corpus. Additionally, we use more sophisticated classifiers to build models for high quality, content-specific article retrieval.

Methods

A. Definitions

In this paper, we chose not to build new criteria to define quality, but instead, we build on existing criteria [7] that the ACP Journal Club uses to evaluate full text articles [15].

The ACP Journal Club is a highly-rated meta-publication. Every month expert clinicians review a broad set of journals [7] in internal medicine, and select articles in these journals according to specific criteria [7] in the content areas of: *treatment, diagnosis, etiology, prognosis, quality improvement, clinical prediction guide,* and *economics*. Selected articles are further subdivided into articles that are summarized and abstracted by the ACP because of their “clinical importance” [15], and those that are only cited because they meet all the quality selection criteria but may not pertain to vitally “important clinical areas” [15]. For the purposes of the present study, abstracted and cited articles published in the ACP Journal Club for a given year are considered high quality and are denoted as ACP+; all other MEDLINE articles not abstracted or cited in the ACP Journal Club, but present in the journals reviewed by the ACP Journal Club, are denoted as ACP-. By using articles abstracted and cited by the ACP Journal Club as our gold standard, we capitalize on an existing, focused quality review that is highly regarded and uses stable explicit quality criteria.

B. Corpus Construction

We constructed two corpora that reflect the progression of our experiments. Corpus 1 has 15,786 MEDLINE records used for high quality treatment and etiology article prediction. Corpus 2 has 34,938 MEDLINE records used for high quality prognosis and diagnosis article prediction. In order to learn high quality models, sufficient ACP+ articles must exist in each category. For our initial experiments including treatment and etiology, we selected a publication time period from July 1998 to August 1999. This chosen period did not yield sufficient ACP+ articles for the prognosis and diagnosis categories so we obtained additional prognostic and diagnostic articles by lengthening the selected publication time period from July 1998 to August 2000. The resulting distribution of positive/ negative articles in each category is 379/15407 in treatment, 205/ 15581 in etiology, 74/ 34864 in prognosis, and 102/ 34836 in diagnosis.

We downloaded all the MEDLINE records in the respective time periods and marked the articles as ACP+. We used a custom script to match word for word the ACP+ title, authors, and journal to the downloaded citations. Next, we downloaded all MEDLINE records from PubMed with abstracts from the journals reviewed by the ACP in the publication period of July 1998 through August 1999 for corpus 1 and July 1998 to August 2000 for corpus 2. Two conditions motivated this period of time. As discussed above, each selected time period provided sufficient ACP+ articles in each category. Selecting a period of several years before the start of the present study gave ample time for the journal club to review the published full text articles for republication in the ACP journal. Thus, to ensure that no ACP+ articles are missed, the ACP journal was reviewed from the *journal* time periods of July 1998 to December 2000 and July 1998 to December 2001 for each respective corpus. From these two selected ACP *journal* time periods, we marked in the *publication* time periods any cited or abstracted articles.

Furthermore, as stated before, we identified 49 journals [7] appearing in the review lists of the table of contents of the first ACP journal in July 1998 to the last ACP journal in December 2001. By collating all articles from these select journal sources that ACP stated it used in preparing the Journal Club, a complete set of references (for the purposes of the current study) was obtained.

At the time of this study, the Esearch and Efetch services of PubMed did not exist [16]. We instead, created custom Python scripts that simulated a user search session to download the MEDLINE records. Each search was limited to the title of one of the 49 journals and set to only retrieve records with abstracts and during the publication period. These MEDLINE records were downloaded in XML format, stored in a MySQL database

[17], and parsed for PubMedID, title, abstract, publication type, originating journal, and MeSH terms with all qualifiers.

C. Corpus Preparation

We partitioned each corpus into *n* fold cross-validation sets to estimate the classification and error of the constructed models. Each cross validation set had a train, validation, and test split with the proportions of ACP+ and ACP- articles maintained in each split.

We chose the number *n* of *n*-fold cross-validation sets based on the frequency of ACP+ high quality articles. For all categories, we chose an *n* of 5. This choice for *n* provided sufficient high quality positive samples for training in each category and provided sufficient article samples for the classifiers to learn the models in our preliminary experiments.

Specifically, the cross-validation sets were constructed as follows. First each corpus was partitioned into 5 disjoint "test" subsets whose union is the complete corpus. For each test split, the remaining 80% of the articles were further partitioned into a 70% "train" split and a 30% "validation" split. In all cases the train, validation, and test splits are chosen so that the proportions of ACP+ articles and ACP- articles are as close as possible to the proportions in the corpus. The validation split was used to optimize any specific learning model parameters. We optimized the models using maximization of area under the ROC curves [18].

D. Article Preparation

The abstracts, titles, and originating journal were parsed into tokens using the algorithm described below and weighted for classifier input. Additionally we extracted MeSH terms including headings and subheadings, and publication types for each MEDLINE record and encoded these as phrases. For example, the publication type *Case Reports* is encoded as a single variable, and following the algorithm below would be encoded as "pt_Case Reports." Next, individual words in the title and abstract were further processed by removal of stop words identified by PubMed [19] such as: "the," "a," "other," etc. that are not likely to add semantic value to the classification. The words were further stemmed by the Porter stemming algorithm which reduced words to their roots [20]. Stemming increases the effective sample by removing word forms often do not add additional semantic value to the classification.

We then encoded each term into a numerical value using log frequency with redundancy (See on-line supplement for mathematical details [7]). The log frequency with redundancy scheme weights words based on their usefulness in making a classification, since words that appear frequently in many articles are assumed to be less helpful in classification than (more selective) words that appear in fewer articles. This weighting scheme was chosen due to its superior classification performance in the text categorization literature [21].

In summary, the algorithm for processing each article is described below:

```
For each article/MEDLINE record in the set
    Extract original journal
    Extract MeSH terms
```

```
        replace all punctuation and spaces with '_'
        associate main headings with each subheading ||i.e. Migraine:etiology and Migraine: therapy||
        precede all terms with 'mh_'  *thus all MeSH terms are encoded as single variables*
    Extract publication types
        precede all terms with 'pt_'
        replace all punctuation with '_'
    For abstract and title words separately
        if title word: precede term with ‘title_’
        convert all words to lowercase
        remove all punctuation and replace with '_'
        remove MEDLINE stop words
        Porter-stem all words
        calculate weights using log frequency with redundancy [21]
        calculate raw frequency occurrence of terms
For each encoded word
    If the word appears in less than 3 documents, remove it from the calculations.
```

Finally, we calculated the raw occurrence of terms in each article. Naïve Bayes and the first version of the Boostexter algorithm are designed to work with discrete data using frequency of term occurrence as input. The second version of Boostexter and support vector machines used the log frequency with redundancy weighted terms as input [22]. In all cases, no term selection was employed, and each algorithm used all available terms for learning.

E. Statistical and Machine Learning Methods

1. Naïve Bayes

Naïve Bayes is a common machine learning method used in text categorization. The Naïve Bayes classifier [23] estimates the probabilities of a class *c* given the raw terms *w* by using the training data to estimate P(*w*|*c*). The class predicted by the Naïve Bayes classifier is the max a-posteriori class.

We coded the algorithm in C as described in Mitchell 1997 [24]. No parameter optimization is necessary for Naïve Bayes. See the online supplement for equations [7].

2. Text-Specific Boosting

Boostexter is a collection of algorithms that apply boosting to text categorization [22]. The idea behind boosting is that many simple and moderately inaccurate classification rules (called the "weak learners") can be combined into a single, highly accurate rule. The simple rules are created sequentially, and for each iteration, rules are created for examples that were more difficult to classify with preceding rules. The prototypical algorithm for boosting is AdaBoost [25]. See the online supplement for mathematical details [7].

The AdaBoost.MR algorithm in the Boostexter suite uses boosted trees to rank outputs with real values. AdaBoost.MR attempts to put correctly labeled articles at the top of the rankings. The algorithm minimizes the number of misordered pairs, i.e. pairs where an incorrectly labeled article is higher in the ranking than a correctly labeled article. The AdaBoost.MR algorithm runs with real valued weights and discrete counts of word frequencies as inputs depending on the version.

3. Support Vector Machines (SVMs)

Support vector machines (SVMs) can function as both linear and non-linear classifiers for discrete and continuous outputs. The type used in this study is the soft margin hyperplane classifier that calculates a separating plane by assigning a cost to misclassified data points. The solution is found by solving a constrained quadratic optimization problem. In addition, for the non-linear case, the problem is solved by using a “kernel” function to map the input space to a "feature" space where the classes are linearly separated. Linear separation in feature space results in a non-linear boundary in the original input space [26-28].

For the text categorization task, the words were weighted using log frequency with redundancy and utilized as features for the linear and polynomial SVMs. We use the soft margin implementation of SVMs in SVM-Light [29]. For the linear SVM, we used misclassification costs of {0.1, 0.2, 0.4, 0.7, 0.9, 1, 5, 10, 20, 100, 1000} for optimization on the validation set. For the polynomial SVM, we used misclassification costs of {0.1, 0.2, 0.4, 0.7, 0.9, 1, 5, 10, 20} and polynomial degrees of {2, 3, 5, 8}. These costs and degrees were chosen based on previous empirical research [30], since the theoretical literature on domain characteristics as it relates to optimal parameter selection is not yet well-developed in this domain. Combinations of both cost and degree were run exhaustively on the validation set, and the optimal cost and degree were applied to the test set in each cross fold validation set. See the on-line supplement for the mathematical details [7].

4. Clinical query filters (CQF)

We ran the category-specific Boolean queries shown in Table II-1 on the corresponding test sets. As described above, two set of Boolean queries exist (i.e. optimized separately for sensitivity and specificity [6]). We measured the optimized sensitivity and specificity values independently for each cross validation set. For the best learning method, we fix these values in each fold and calculate the corresponding sensitivity, specificity, and precision. We report the average optimized and matched values across all folds in Table II-2.

F. Evaluation Criteria

We used 4 evaluation criteria: (a) area under the receiver operating curve (ROC) (AUC) of each method with statistical comparison between methods using the Delong paired ROC comparison test [31], (b) 11 point precision-recall curves, (c) comparison to the specificity of the clinical query filters at the point of equal sensitivity, and (d) comparison to the sensitivity of the clinical query filters at the point of equal specificity. For (c) and (d), we used McNemar's test to statistically compare each method to the best learning method.

We calculate the AUC and ROC for each method in each fold, and calculate the averaged statistical significance of the difference of the best performing method over all folds to each of the other methods using Delong method [31]. For a single learning method, we estimate the statistical significance across all cross validation sets. We averaged the p-values for all the sets to obtain an empirical mean. We statistically evaluate this empirical mean by examining the distribution of means obtained by randomly permuting a complete experiment (i.e. in this case, randomly permuting 5 cross validation sets for one method and obtaining a permuted mean) 500 times. With the empirical mean and the distribution of means created by the permutations, we report a significance value for the empirical mean and thus conclude a statistical p-value difference between the best learning method and the compared method.

Note that although several parametric tests for comparing mean p-values exist, they assume independence between measurements [32]. These independence assumptions do not apply in an *n*-fold cross validation setting; thus we resort to a random permutation test here.

We compare the sensitivity and specificity of the machine learning methods to the sensitivity and specificity of the respective optimized Boolean clinical query filter. The query filters returned articles with the query terms present whereas the learning algorithms return a score. To make the comparison, in each fold, we fixed the sensitivity value returned by the sensitivity-optimized filter, and varied the threshold for the scored articles until the sensitivity was matched. We report the averaged fixed sensitivity and matched threshold in Table II-2. The same procedure was run for the specificity returned by the optimized specificity filter.

We assessed the statistical significance of differences of sensitivities (or specificities) between the best learning method and the clinical query filter Boolean models using McNemar's test (calculated for each cross-validation set) [33]. To report the significance across all cross validation sets, we followed the same procedure as described above in comparing ROC curves. Instead of using the Delong method, we instead compare the best learning method to the Boolean models with McNemar’s test for all the sets to obtain an empirical mean. We statistically evaluate this empirical mean by examining the distribution of means obtained by randomly permuting a complete experiment (i.e. in this case, randomly permuting 5 cross validation sets and obtaining a permuted mean) 500 times.

Table II-2 - Best learning method compared to clinical query filters fixed at optimal sensitivity and specificity. The first number is the average across 5 folds. The numbers in parenthesis report the minimum and maximum value across the 5 folds. Cells in bold denote the performance for the filter optimized for sensitivity or specificity respectively.

Category	Optimized for	Method	Sensitivity	Specificity	Precision
Treatment	Sensitivity	Query Filters	**0.96(0.91-0.99)**	0.75(0.74-0.76)	0.09(0.08-0.09)
		Poly SVM		0.86(0.68-0.93)	0.18(0.07-0.25)
	Specificity	Query Filters	0.4 (0.37-0.42)	**0.96 (0.95-0.96)**	0.19 (0.17-0.21)
		Poly SVM	0.80(0.74-0.83)		0.33(0.31-0.34)
Etiology	Sensitivity	Query Filters	**0.70 (0.61-0.78)**	0.85 (0.85-0.86)	0.06 (0.06-0.06)
		Poly SVM		0.95(0.92-0.97)	0.15(0.11-0.21)
	Specificity	Query Filters	0.28(0.24-0.37)	**0.93 (0.92-0.94)**	0.05 (0.04-0.06)
		Poly SVM	0.76(0.68-0.78)		0.12(0.12-0.12)
Prognosis	Sensitivity	Query Filters	**0.88 (0.80-0.93)**	0.70 (0.70-0.71)	0.006 (0.006-0.007)
		Poly SVM		0.71 (0.32-0.86)	0.009 (0.003-0.013)
	Specificity	Query Filters	0.51 (0.33-0.80)	**0.94 (0.94-0.94)**	0.02 (0.011-0.026)
		Poly SVM	0.62(0.60-0.67)		0.02(0.02-0.02)
Diagnosis	Sensitivity	Query Filters	**0.95(0.86-1.0)**	0.7(0.69-71)	0.009 (0.009-0.010)
		Poly SVM		0.53 (0.04-0.95)	0.015(0.003-0.048)
	Specificity	Query Filters	0.67(0.48-0.80)	**0.96 (0.96-0.96)**	0.048 (0.034-0.056)
		Poly SVM	0.77(0.70-0.86)		0.055(0.049-0.059)

Results

A. Area under the receiver operating curve analysis

The areas under the receiver operating characteristic curves (AUC) for each category averaged over 5 folds are presented in Table II-3. Values upwards of 0.91 with ranges for the best learning methods suggest that the corresponding learning methods can distinguish very well between positive and negative class articles. The polynomial SVM turned out best, and it was compared as a baseline to all other learning methods and the clinical query filters. In the treatment and etiology categories, in nearly all cases except Boostexter raw in etiology, the difference of the polynomial SVM output to the other methods was not due to chance. In contrast, in the sample limited diagnosis category, the difference between the polynomial SVM output and the Boostexter algorithms and the linear SVM may be due to chance. Similarly, in the sample limited prognosis category, the linear and polynomial SVM difference may be due to chance as well.

The ROC curves for each category and learning method are depicted in Figure II-1. In all cases, the learning methods perform well with the exception of Naïve Bayes in the prognosis and diagnosis categories. Finally, in each ROC graph, the corresponding clinical query filter performances are shown by small X's. The leftmost symbol corresponds to fixed specificity and the rightmost symbol corresponds to fixed sensitivity.

B. 11 point precision recall

We further compare qualitatively the clinical query filters to the best learning method (polynomial SVM) in each category in Figure II-2. For each category, we mark on the 11 point precision recall graph the corresponding precision recall performance for the optimized sensitivity and specificity clinical query filters. The leftmost point is the filter optimized for specificity and the rightmost point is the filter optimized for sensitivity. For treatment, etiology,

Table II-3 - Area under the Receiver Operating Curve (AUC) Performance of each machine learning method in each category.

Diagnosis						**Prognosis**				
Learning method	Average AUC*	Min AUC*	Max AUC*	^Significance (Delong)		Learning method	Average AUC*	Min AUC*	Max AUC*	Significance (Delong)
Naïve Bayes	0.82	0.80	0.84	0.001 (0)		Naïve Bayes	0.58	0.47	0.66	0 (0)
Boostexter –Weighted	0.87	0.85	0.90	0.10 (0)		Boostexter – Weighted	0.71	0.56	0.86	0.01 (0)
Boostexter – Raw Frequency	0.94	0.91	0.97	0.43 (0.03)		Boostexter – Raw Frequency	0.79	0.73	0.85	0.04 (0)
Linear SVM	0.95	0.93	0.97	0.11 (0)		Linear SVM	0.91	0.86	0.94	0.39 (0.01)
Polynomial SVM	**0.96**	**0.95**	**0.98**	**N/A**		**Polynomial SVM**	**0.91**	**0.87**	**0.95**	**N/A**
Treatment						**Etiology**				
MLmethod	Average AUC	Min AUC	Max AUC	Significance (Delong)		MLmethod	Average AUC	Min AUC	Max AUC	Significance (Delong)
Naïve Bayes	0.95	0.94	0.95	0.01 (0)		Naïve Bayes	0.86	0.84	0.88	0.02 (0)
Boostexter –Weighted	0.94	0.92	0.95	0.03 (0)		Boostexter – Weighted	0.85	0.83	0.87	0.01 (0)
Boostexter – Raw Frequency	0.94	0.93	0.96	0.01 (0)		Boostexter – Raw Frequency	0.90	0.88	0.93	0.25 (0)
Linear SVM	0.96	0.95	0.97	0.03 (0)		Linear SVM	0.91	0.86	0.93	0.03 (0)
Polynomial SVM	**0.97**	**0.96**	**0.98**	**N/A**		**Polynomial SVM**	**0.94**	**0.89**	**0.95**	**N/A**

* - average, minimum, and maximum AUC across the respective number of folds for each category.

^ - Mean significance using the Delong method across all folds comparing the best learning method in each category (Polynomial SVM) with each other learning method. The number in parenthesis is the significance produced by random permutation test as described in Section IIIF

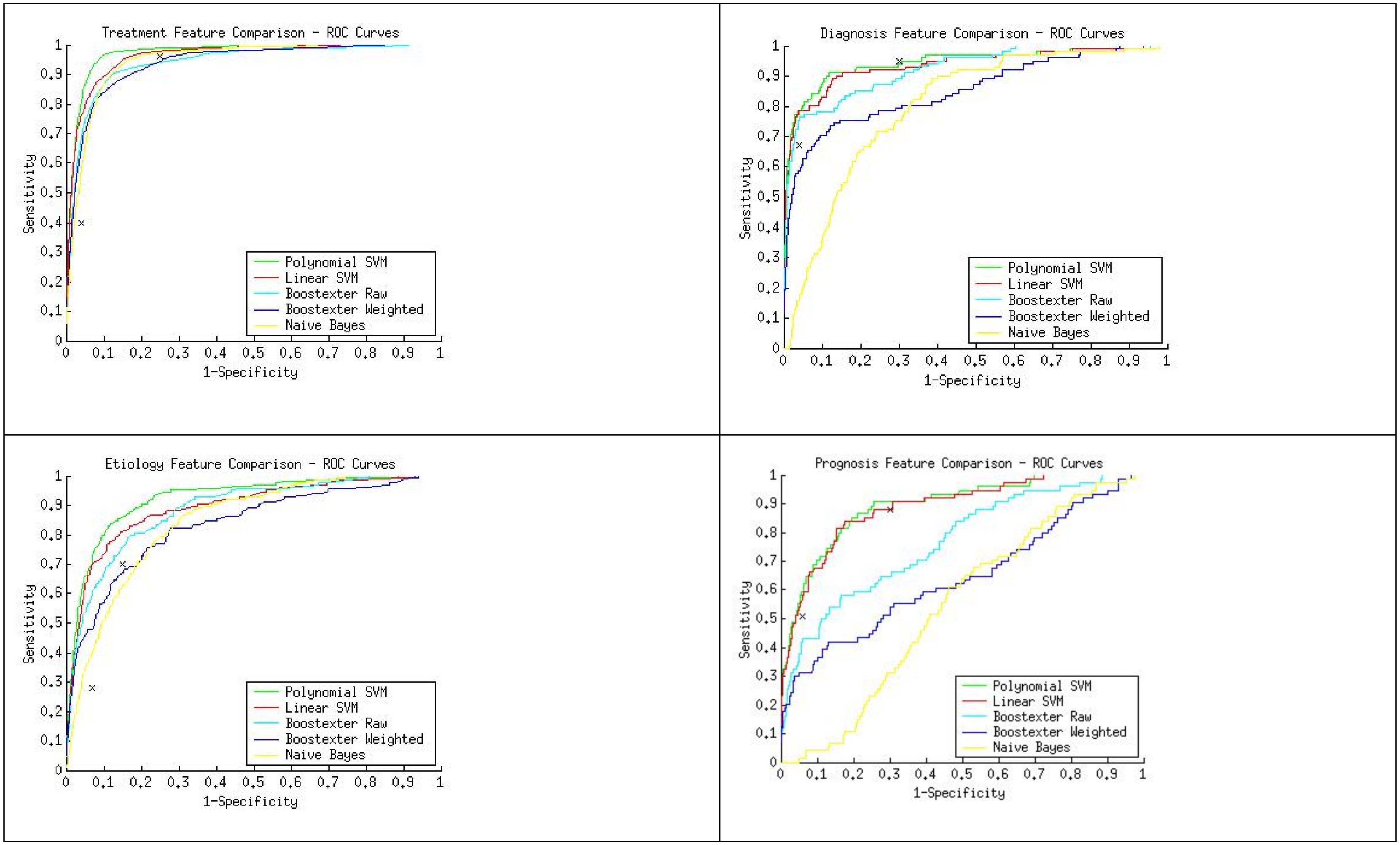


Figure II-1 - Receiver Operating Curves for Each Category
x's – clinical query filter performance at optimized sensitivity (right x) and specificity (left x)

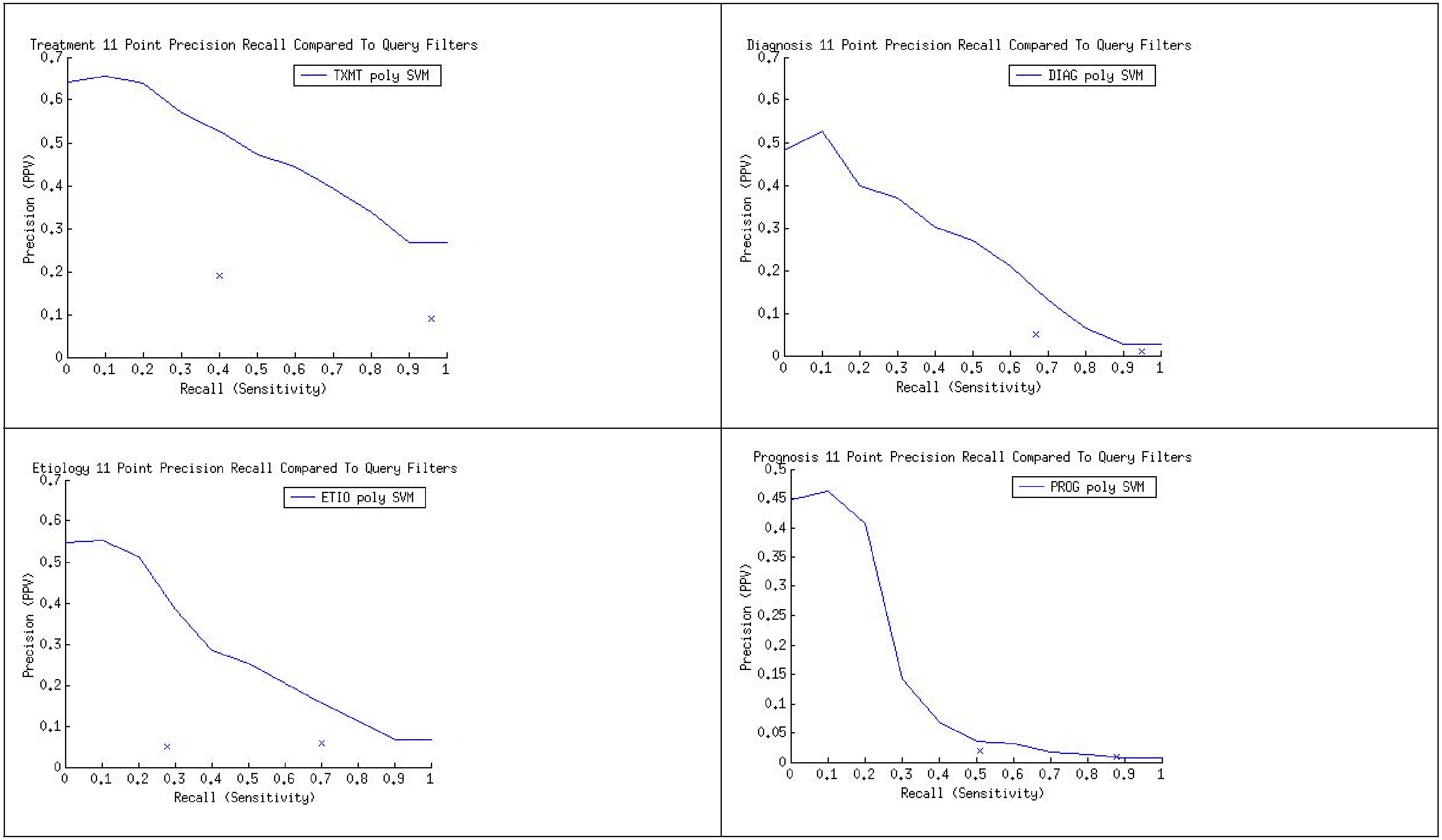


Figure II-2 – 11 Point Precision Recall curves compared to optimized sensitivity and specificity clinical query filters.
plus, square, x's, triangles – clinical query filter optimized for sensitivity (left mark) and specificity (right mark)

and diagnosis, the polynomial SVM performed better than either optimized clinical query filter using this metric. For prognosis, the polynomial SVM performed as well as the clinical query filters using this metric.

C. Comparison to clinical query filters

For the most part, the learning methods outperformed the query filters for each sensitivity, specificity, and precision measure. Table II-2 compares the *best* learning method by AUC and the results of the clinical query filters fixed for sensitivity and specificity respectively for each category. The average with the ranges across 5 folds across all cross-validation sets appear inside parentheses.

In comparison to the clinical query filters, the polynomial SVM has better performance in the treatment and etiology categories. In the prognosis category the polynomial SVM model and the clinical query filters perform similarly. In the diagnosis category, the polynomial SVM performs better than the specificity optimized filter but worse than the sensitivity optimized filter (See discussion). The polynomial SVM model for treatment and etiology at a threshold that matches the sensitivity of the sensitivity-optimized clinical query filter has at least double precision compared to the clinical query filters though remaining below 20% in both categories. Specificity of the polynomial SVM model is also better (by approximately 10% in both categories). Likewise, in the same categories, the polynomial SVM model at a threshold that matches the specificity of the specificity-optimized clinical query filter has almost double precision compared to the clinical query filters. Sensitivity of the polynomial SVM model is also better (by 40% and 48% respectively). For the prognosis category, the polynomial SVM model performs comparably to the sensitivity and specificity-optimized clinical query filters. For diagnosis, the polynomial SVM model has a 10% improvement in sensitivity for the specificity optimized filter, but a 17% decline in specificity for the sensitivity optimized filter (See discussion for details).

Table II-4 – McNemar’s test p-values averaged over 5 folds with significance tests. The permutation significance is produced by random permutation tests as described in Section IIIF.

Category	Filter Compared	Mean p-values	Permutation Significance
Treatment	Sensitivity	< 0.0001	< 0.0001
	Specificity	0.019	< 0.0001
Etiology	Sensitivity	< 0.0001	< 0.0001
	Specificity	0.34	0.14
Prognosis	Sensitivity	< 0.0001	< 0.0001
	Specificity	0.95	1.0
Diagnosis	Sensitivity	0.07	< 0.0001
	Specificity	0.90	1.0

Table II-4 compares statistically the polynomial SVM and the clinical query filters using McNemar’s test. As described in the methods section, we report (a) the average p-values across all cross validation sets and (b) the significance using a random permutation test.

When comparing the optimized *sensitivity* filters to the polynomial SVM, the mean p-values are significant at the 0.05 level except for the sensitivity optimized diagnosis filter at the 0.07 level. Thus, the improvements compared to the clinical query filters in both precision and specificity are not due to chance.

When comparing the optimized *specificity* filter to the polynomial SVM, in etiology, prognosis, and diagnosis categories, the mean p-values are *not* significant whereas in the treatment category, the polynomial SVM models *are* significant at the 0.05 level. Hence we conclude that the differences between the polynomial SVM fixed at optimized specificity and the query filters are not due to chance in the treatment category, but may be due to chance in the other 3 categories. We speculate that in these 3 categories, non-significant differences are due to the low ratios of ACP+ to ACP- articles (i.e. low priors).

Discussion

We have shown that machine learning methods applied to categorizing high quality articles in internal medicine for a given year perform better than currently used Boolean methods in most categories. This work is a step toward efficient high-quality article retrieval in medicine.

A. Performance in the Diagnosis Category

In light of the comparable or superior performance of the SVM model over the clinical query filters in treatment, etiology, and prognosis, the lower performance of the diagnosis polynomial SVM versus the sensitivity-optimized query filter warranted further attention.

Recall that we match the sensitivity returned from the optimized diagnosis Boolean query to the sensitivity produced by varying the threshold for the SVM output. Because the number of positive articles in the diagnosis category is very small (and even smaller within the splits of cross-validation), and because the Clinical Query Filters exhibit very high sensitivity in the content category, even a small number of outliers (i.e., MEDLINE documents receiving a low score) in terms of SVM model scores, will result in significant reduction of the specificity once we set the SVM threshold to match the near-perfect sensitivity of the Clinical Filters.

Indeed, we identified such outliers and verified that they were the source of the reduced performance in the diagnosis category once we fix the thresholds to match the CQF sensitivity. By close examination we found that the ACP+ articles scored low because the terms used to identify these articles were not used in training of the SVM model. More specifically, MeSH subheadings were not encoded individually. For example, one of the ACP+ articles scoring low was identified by the diagnosis clinical query filters with the MeSH subheading “diagnosis” (See Table II-1). Recall from the article preparation procedure in Section IIID that mesh *subheadings* are not encoded explicitly, but only as part of the matching major heading. Thus “diagnosis” would not be encoded individually, but only as part of the major heading as in “Migraine:diagnosis.”

If the ACP+ article was encoded as "Pneumonia:diagnosis," it would not score high. The SVM classifiers did not have sufficient information to give some ACP+ articles a high score, since none of these words were found in the text.

It is evident that this problem can be fixed simply by encoding the subheadings individually in future versions of the models discussed here. However we do note that in such circumstances, using the human Mesh indexing provides a slight edge over not using them.

B. Implicit Selection Bias

A potential drawback of the constructed models is that they may reflect implicit selection biases by the editors of the ACP Journal Club, and the high quality articles selected by the models are not based on sound methodology. For example, it is conceivable that editors for a particular year could have a favourable bias toward a particular subject, and the subject rather than the methodology causes a high quality classification.

We answer this concern through cross-validation and a method presented in [34] to convert the models to Boolean queries. Specifically, we built Boolean models using an approximate Markov Blanket feature selection technique [34] modified from [35] to obtain the set of minimal terms, and a decision tree to build the corresponding Boolean query. The feature selection/ decision tree method presented in [7] shows that the models emphasize methodological words in nature rather than topic specific ones.

C. Labor Reduction

The machine learning based methods may significantly reduce labor through automated term selection, reliance on an existing, publisher-based, expert derived gold standard, and a reduced feature set without manually assigned MeSH terms and publication types that has equivalent performance to the full set with terms and types.

Recall from the background section the strategy for development of the clinical query filters [6]. In the Haynes approach, significant time is spent interviewing people for the selected terms, building the gold standard, and running an exhaustive search through the space of term disjunctions. In addition, the filters rely on MeSH terms and publication types that must be assigned before the filters can be used.

In contrast, the methods here are less labor intensive. First, there is no selection of terms as these are implicit in the training articles. Second, we have a framework for automatic generation of a gold standard through the ACP Journal Club that is reliable and reproducible. Manual review is not needed as long as the ACP journal is electronically available. Finally, we use sophisticated classifiers that can build models in 4-8 hours (depending on model and experiment design) on a Pentium 4, 2GHz with full term sets versus several days depending on the number of selected terms with the exhaustive search of term disjunctions [6].

In an additional experiment, we compared the inclusion/ exclusion of manually assigned, labor-intensive MeSH terms and publication types (NLM assigned terms) as model input features. We compared the ROC performance of a feature set inclusive NLM assigned terms to a feature set without both. The ROC curves in Figure II-3 show that the reduced feature set without NLM assigned terms has comparable ROC curves to the feature set inclusive of these terms. Though we do not show the results here (see on-line supplement for further details [7]), each average AUC was *not* significantly different for each feature set using the Delong method [31]. The results suggested, with our methods, we can make quality and content determinations without the labor-intensive NLM term indexing process. Note that we do not advocate abandoning human indexing in general, but for this task, no additional benefit is gained from manual term assignments.

D. Extensions

Another avenue to explore is the use of additional predictor information. For example, we hypothesize that additional information such as general word location, impact factors, citation information, author locations, or user feedback information may improve model performance.

We also plan to extend these models to areas outside of internal medicine. One approach is to build a gold standard that considers articles in other specialties. *Evidence Based Medicine* is the sister journal of the *ACP journal* that could be used for a more general gold standard, since its scope of review covers all aspects of medicine.

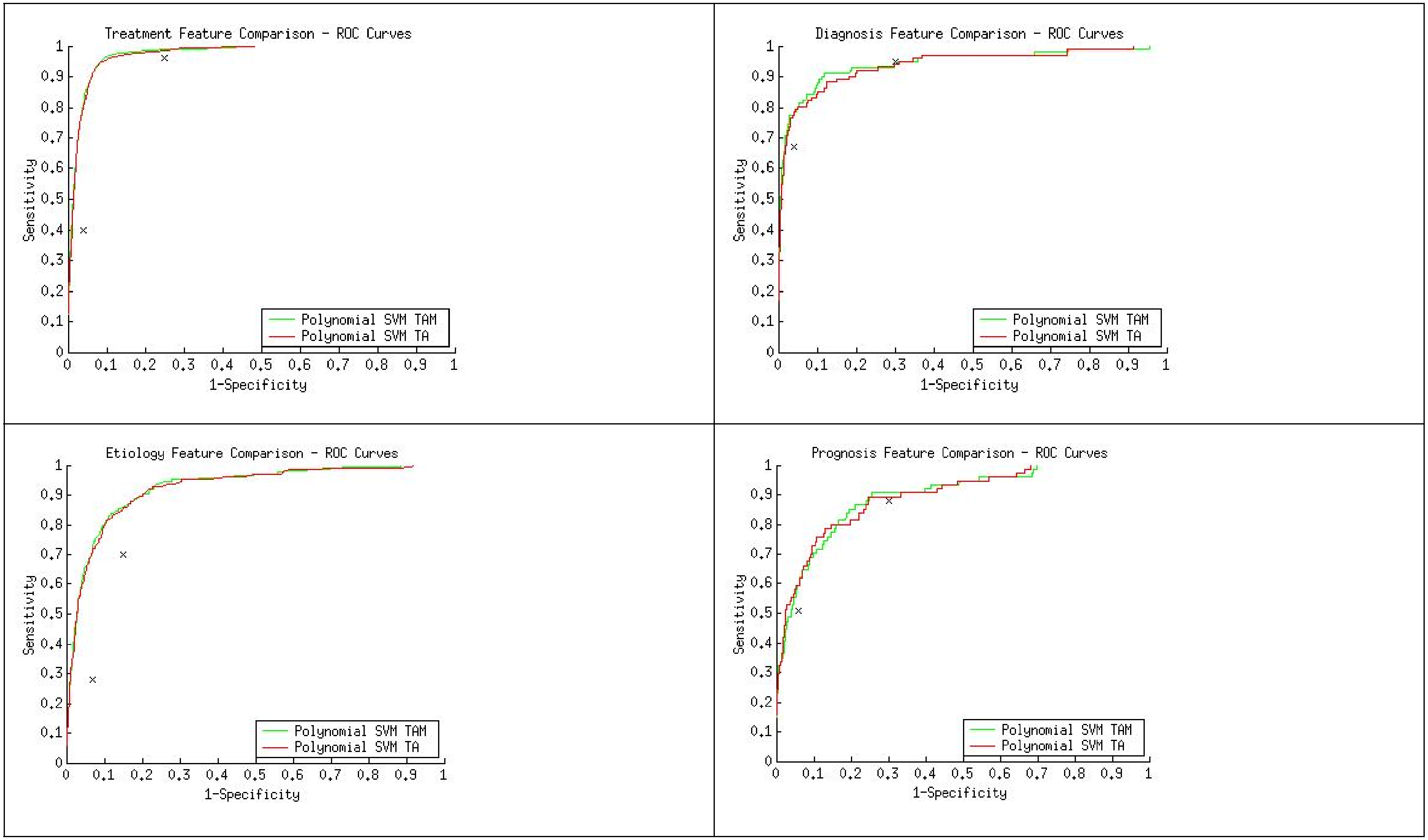


Figure II-3 – Title+Abstract (TA) vs. Title+Abstract+MeSH+Publication Types (TAM) Performance Comparisons.
x's – clinical query filter performance at optimized sensitivity (right x) and specificity (left x)

E. Limitations

In general, the prognosis and diagnosis samples sizes are limited. We chose not to alter the ratio of positive to negative articles to maintain the priors across all learning tasks and produce realistic estimates of future performance. The small priors for both these categories make learning difficult. Nevertheless, with these sample sizes, our system performs at least comparably to the clinical query filters in prognosis and in some cases in diagnosis.

Another admitted limitation of our comparisons to the clinical query filters is that the new models and filters were built for the exact same goals but with different gold standards. Our comparisons simply show that the new models implement the present gold standard better than the clinical query filters. In the future, using an independent gold standard and evaluating both methods trained on independent sets would strengthen this comparison.

A potential limitation of any information retrieval study is the choice of gold standard. A gold standard is only as good as the experts brought together to create it. The use of the ACP Journal Club articles meets our criteria, and we propose that currently, is the best general method to create such gold standards. The journal club articles are easily obtained from their website, the cited articles are readily available for use by other researchers, and the gold standard is created by recognized experts and editors in the field of internal medicine.

This work is a step toward more efficiently returning high quality articles. The work does not address explicitly the utility of these models *in a clinical setting* or outside of internal medicine. Finally, the learning method's built models are constrained to one specific time period in internal medicine.

Summary

Text categorization methods can learn models that identify high quality articles in specific content areas (etiology, treatment, diagnosis, and prognosis) by analyzing MEDLINE records in internal medicine using the operational gold standard of articles that match the ACP inclusion criterion for methodologic rigor. These learning methods exhibit high discriminatory performance as measured by the AUC. The performances are also comparable or better than the Boolean based clinical query filters for each category by direct comparisons of sensitivity, specificity, and precision at fixed levels and by 11 point precision recall comparisons. Polynomial SVMs have the best performance while linear SVMs came close in terms of AUC. We presented an efficient and improved means for identifying high quality articles in internal medicine.

Acknowledgements

We acknowledge the Vanderbilt MSTP program and NLM grant LM007948-02 for support. We also acknowledge Dr. Randolph Miller, the anonymous reviewers, and the senior editor for their valuable comments and suggestions.

References

[1] Bigby M. Evidence-based medicine in a nutshell. Arch Dermatol. 1998 Dec;123(12):1609-18.

[2] Sackett DL, Richardson WS, Rosenberg W, Haynes BR. Evidence Based Medicine: How To Practice and Teach EBM. Edinburgh: Churchill Livingstone 1998.

[3] Library CC. [cited; Available from: http://www.cochrane.org

[4] Medicine EB. [cited; Available from: http://ebm.bmjjournals.com

[5] Journal A. [cited; Available from: http://www.acpjc.org

[6] Haynes B, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing Optimal Search Strategies for Detecting Sound Clinical Studies in MEDLINE. JAMIA. 1994 November/December;1(6):447-58.

[7] Aphinyanaphongs Y, Aliferis CF, Tsamardinos I, Statnikov A, Harding D, Miller RA. On-line Supplement to Text Categorization Models For Retrieval of High Quality Articles in Internal Medicine. http://stagingmcvanderbiltedu/discover/public/supplements/TextCat/. 2004.

[8] PubMed. [cited; Available from: http://www.ncbi.nlm.nih.gov/PubMed/

[9] Wilczynski N, Haynes B. Developing Optimal Search Strategies for Detecting Clinically Sound Causation Studies in MEDLINE. Proceedings AMIA Symposium; 2003; Washington DC; 2003. p. 719-23.

[10] Wong S, Wilczynski N, Haynes R, Ramkissoonsingh R. Developing Optimal Search Strategies for Detecting Sound Clinical Prediction Studies in MEDLINE. Proceedings of AMIA Symposium; 2003; Washington DC; 2003. p. 728-32.

[11] Robinson KA, Dickersin K. Development of highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed. International Epidemiological Association. 2002;31:150-3.

[12] Nwosu C, Khan K, Chien P. A Two-Term MEDLINE Search Strategy for Identifying Randomized Trials in Obstetrics and Gynecology. Obstetrics and Gynecology. 1998 April;91(4).

[13] Shojania KG, Bero LA. Taking Advantage of the Explosion of Systematic Reviews: An Efficient MEDLINE Search Strategy. Effective Clinical Practice. 2001 July/August;4(4):157-9.

[14] Bachmann LM, Coray R, Estermann P, Reit GT. Identifying Diagnostic Studies in MEDLINE: Reducing the Number Needed to Read. JAMIA. 2002 Nov/Dec;9(6):653-8.

[15] Purpose and Procedure. ACP Journal. 1999 July/August;131(1):A-15 - A-6.

[16] E-Utilities. [cited; Available from: http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

[17] MySQL. [cited; Available from: http://www.mysql.com/

[18] Centor RM. The Use of ROC Curves and Their Analyses. Medical Decision Making. 1991 April - June;11(2):102-6.

[19] Stopwords M. [cited; Available from: http://www.princeton.edu/~biolib/instruct/MedSW.html

[20] Porter MF. An algorithm for suffix stripping. Program. 1980;14(3):130-7.

[21] Leopold E, Kindermann J. Text Categorization with Support Vector Machines. How to Represent Texts In Input Space? Machine Learning. 2002;46:423-44.

[22] Schapire RE, Singer Y. Boostexter: A Boosting-based System for Text Categorization. Machine Learning. 2000;39(2/3):135-68.

[23] Joachims T. A probabilistic analysis of the Rocchio Algorithm With TFIDF for text categorization. 14th International Conference on Machine Learning; 1997; Nashville, TN: Morgan Kauffman; 1997. p. 143-51.

[24] Mitchell TM. Machine learning. New York: McGraw-Hill 1997.

[25] Schapire RE. Theoretical views of boosting and applications. Tenth International Conference on Algorithmic Learning Theory; 1999; 1999.

[26] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and other Kernel-based Learning Methods: Cambridge University Press 2000.

[27] Burges C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 1998;2:121-67.

[28] Vapnik V. Statistical Learning Theory: Wiley 1998.

[29] Joachims T, ed. Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning.: MIT-Press 1999.

[30] Aphinyanaphongs Y, Aliferis CF. Text Categorization Models for Retrieval of High Quality Articles in Internal Medicine. Proceedings AMIA Symposium; 2003; Washington DC; 2003.

[31] Delong E, Delong D, Clarke-Pearson D. Comparing the Area under Two or More Correlated Receiver Operating Characterisitic Curves: A Nonparametric Approach. Biometrics. 1988 Sept;44(3):837-45.

[32] Cooper H, Hedges L. The Handbook of Research Synthesis: Russell Sage Foundation Publications 1994.

[33] Pagano M, al e. Principles of Biostatistics: Duxbury Thompson Learning 2000.

[34] Aphinyanaphongs Y, Aliferis CF. Learning Boolean Queries for Article Quality Filtering. MEDINFO; 2004; San Francisco, CA; 2004.

[35] Aliferis C, Tsamardinos I, Statnikov A. HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection. Proceedings AMIA Symposium; 2003; Washington DC.; 2003.

[36] Stopwords M. [cited 11-26-2006]; Available from: http://biolib.princeton.edu/instruct/MedSW.html

[37] MeshBrowser. [cited; Available from: http://www.nlm.nih.gov/mesh/MBrowser.html

Learning Boolean Queries for Article Quality Filtering

Aphinyanaphongs Y, Aliferis C. "Learning Boolean Queries for Article Quality Filtering." In: MEDINFO; 2004; San Francisco, CA; 2004.

Abstract

Prior research has shown that Support Vector Machine models have the ability to identify high quality content-specific articles in the domain of internal medicine. These models, though powerful, cannot be used in Boolean search engines nor can the content of the models be verified via human inspection. In this paper, we use decision trees combined with several feature selection methods to generate Boolean query filters for the same domain and task. The resulting trees are generated automatically and exhibit high performance. The trees are understandable, manageable, and able to be validated by humans. The subsequent Boolean queries are sensible and can be readily used as filters by Boolean search engines.

Introduction

The pace of research far overcomes the ability of modern health professionals to be up to date about all the recent research developments and current best practices. By one account, a general physician reviewing just 20 clinical journals in adult internal medicine would have to read 19 articles a day for 365 days a year to keep up [1]. Increasingly, physicians are turning to electronic sources for their information needs. Services like

MDConsult [2], Up2Date [3], and Pubmed Central [4] evaluate and abstract research articles.

However, the final authority on what constitutes best medical practices and high quality knowledge is provided by the primary sources themselves (i.e the biomedical research literature). Thus there exists a great need for a way to identify the most important of the primary sources, that is the original research, the methodological quality and scope of which are likely to yield the highest benefit to the healthcare professionals.

A primary point of practical significance is the technology and overall process for constructing quality filters to return these primary sources. More specifically, in most cases, filters consist of Boolean queries that were formulated by taking human-derived queries and modifying them, or by stringing together words that are deemed intuitive by human experts for some domain in disjunctions or conjunctions and evaluating their performance [5, 6]. A more structured, yet still, ad-hoc approach was taken to generate Boolean queries to return high quality content related articles in [7] and [8]. In the pioneering study in [8], experts were polled, and words that were deemed relevant to a content area were selected. The exact combination of words was optimized separately for sensitivity and specificity by a brute force search of all disjunctions of the selected words (up to a small number of words per query). The resulting queries perform well, and are featured in the clinical queries (CQF) link in PubMed [4]. Alternately in [7], word frequencies in the abstract were used to identify candidate terms. These terms are individually evaluated for sensitivity and precision, and the terms with the highest (sensitivity * precision) product were combined in a disjunctive Boolean query to find diagnostic studies. The authors report improved performance over the CQF diagnostic filter.

The authors of the present article in [9] address the problem of returning quality articles by running a suite of powerful classifiers on a suitable corpus and not rely on human experts. While the resulting models perform very well, a question remains as to (a) their understandability by humans, and (b) their usability through Boolean based systems such as Pubmed. Even though the Boolean model can capture any set of documents, the process of formulating such queries, especially by humans, can be challenging. Indeed, analysis of search engine logs show that most search engine users avoid Boolean formulations [10].

Flake et. al. recently introduced a hybrid approach that converts corpus-based SVM models to Boolean queries in the web domain [11]. Their method combines in an ad-hoc manner a linear approximation to a polynomial SVM classifier with a modified Adaboost [12] algorithm to convert the original polynomial SVM models to sets of Boolean Queries (also referred to as “query modifications” in the information retrieval literature). The Flake et al. method is highly heuristic and not guaranteed to perform well in specific data and problem domains, however.

Thus the motivation for this paper is how to convert sophisticated machine learning models into usable queries. The application of SVMs to current information retrieval systems is not straightforward and would require a dedicated system built expressly for this purpose. To bridge the gap and give users applicable technology, we explore the formulation of Boolean queries from a training corpus that includes examples of the high quality content specific articles.

Specifically, we ask the question:

Is it possible to automatically construct Boolean queries from a corpus using machine learning techniques such that the Boolean queries have as good classification performance as the SVM models, and are the resulting Boolean queries human-readable, manageable, and simple for use in current search engines?

Throughout the present paper, we use “word”, “term”, “feature”, and “variable” interchangeably. The choice of word depends on the appropriate context in which it is found.

Methods

Corpus Preparation

We use for the present study a modified version of the corpus in [9]. This corpus uses the ACP journal as a gold standard for both content and quality of articles [13]. The ACP journal is a meta-publication that routinely reviews over a hundred journals for articles that meet its selection criteria. Articles that are abstracted or cited by the ACP are considered positive instances and all other articles in the same journals to be negative. A more detailed description of the gold standard construction methodology can be found in [9]. The criteria for inclusion in ACPJ can be found in [13].

We selected the treatment content area for several reasons. This area had sufficient sample to represent the concepts for a high-quality treatment article. The criteria for selection are simple, and the predominant class of questions asked by physicians is treatment related [14].

The conversion of documents to a format suitable for the machine learning algorithms followed the procedures in [9] closely. The articles in the ACP selected journals were cross-referenced in PubMed, and the title, abstract, and MeSH terms parsed. The processing of the terms differs from [9] in that title and abstract terms were represented separately rather than as one group.

The resulting terms were encoded as binary variables (either appearing in the document or not) in all documents. The final treatment category counts included 397 positive documents and 15407 negative documents with 27891 unique words.

The articles were further split into a training, validation, and test set, with 221 positive / 8998 negative, 76 positive/ 3081 negative, and 82 positive/ 3328 negative documents respectively. A single split was selected because the sample size was large enough, and utilizing a single split simplified the creation of a single Boolean query by removing concerns about how to combine the queries from each split.

Support Vector Machine Classifiers

We used a support vector machine (SVM) from our previous experiments as an empirical "upper bound" on the performance of the binary encoded test set. SVMs function as both linear and non-linear classifiers. They maximize the margin between the instances belonging to different classes. The solution that generalizes best to unseen instances is found by solving a constrained quadratic optimization problem in terms of the patterns that lie on the margin (i.e. support vectors) [15].

We use a Matlab [16] wrapper [17] for Thorsten Joachim's SVM-light [18]. This implementation utilizes a decomposition method to make learning a large number of examples tractable [19]. We use misclassification costs of {0.1, 0.2, 0.4, 0.7, 1.0, 2.0} and degrees of {1, 2, 5} on the validation sets. The best performance combination of degree and cost was used on the test set.

Decision Tree Classifiers

Our primary means to generate Boolean queries is induction of decision trees. The reason for this choice is that the output of a decision tree maps well to Boolean queries. Each leaf of the decision tree corresponds to a path that describes the conjunction of word absence or presence for a classification.

In the text categorization domain, decision trees are a learning method that attempts to partition a training set based on individual words that describe the domain. The extensive work of Apte and Weiss [20], demonstrated that decision trees can produce superior classification performance in text while producing trees that are understandable. Our work extends the findings of Apte in several ways. First, we construct and apply the work to a new task. Second, we introduce new feature selection methods. Third, we analyze the trees in this problem domain to address the understandability and manageability of the resulting queries.

In this paper, we use the CART implementation of decision trees in Release 13 of Matlab with the gini index of diversity [21] to rank the relevant features. The full tree is pruned based on retaining a performance of at least 1% of the maximum performance on the validation set with the smallest tree size. For example, suppose the best tree performs at 92% AUC with 10 nodes and a smaller tree performs at 91% AUC with 5 nodes. We would select the smaller tree as it retains at least 1% AUC of the maximum.

The Flake algorithm was implemented by the first author in Matlab following the description in [11] since public domain code is not currently available.

Feature Selection Algorithms

Decision trees are known to suffer from the curse of dimensionality [22]. As the number of features increases, the increase in sample size must grow exponentially in the worst case, or the decision tree will not generalize well. To overcome this problem, we use several feature selection algorithms with the decision tree.

In our first evaluation of the method we employ three variable selection algorithms:

Linear and Approximate-Polynomial Recursive Feature Elimination (RFE_L RFE_{PA})

RFE builds on the power of SVM classification. The basic procedure can be summarized as follows [23]:

1. Build an SVM classifier using all ***V*** features

2. Compute weights of all features and choose the first |***V***|**k* features (sorted by weight in decreasing order, *k* being a feature set cardinality reduction parameter, typically set to 0.5)

3. Repeat steps 1 and 2 until an empty feature set is produced

4. Choose among all feature subsets created the one that gives the best performance in a validation set

Linear RFE (RFE_L) uses linear SVMs in step 1 as the name implies. In step two features are selected by their weights. In Approximate-Polynomial RFE (RFE_{PA}) a polynomial-kernel SVM is used in step 1 while Step 2 uses, instead of weights, ranking coefficients such that the ranking coefficient of the feature ***i*** is the change of cost function by removing feature ***i***. As a speed-up heuristic, one does not recompute Lagrange coefficients while ranking features. We also note that in the linear case, non-linear RFE is identical to the linear RFE. The exact mathematical formulations and parameter values used for both methods can be found in [23].

HITON-PC_{FW} (filtered and wrapped HITON PC)

HITON is a feature selection algorithm introduced in [24] that combines induction of Markov Blankets and wrapping (i.e., heuristic search over variables subsets) to identify the smallest variable subset that gives optimal classification performance. It was shown by its authors (a) to be sound given the distributional assumption of faithfulness, universal approximator learners, and a quadratic loss misclassification function (for details please see the original publication); and (b) to have superior variable reduction performance (while maintaining optimal or near-optimal classification performance) to a range of state-of-the-art variable selection methods across a representative sample of biomedical tasks, including text categorization. Given HITON's powerful reduction capabilities we apply it in our experiments.

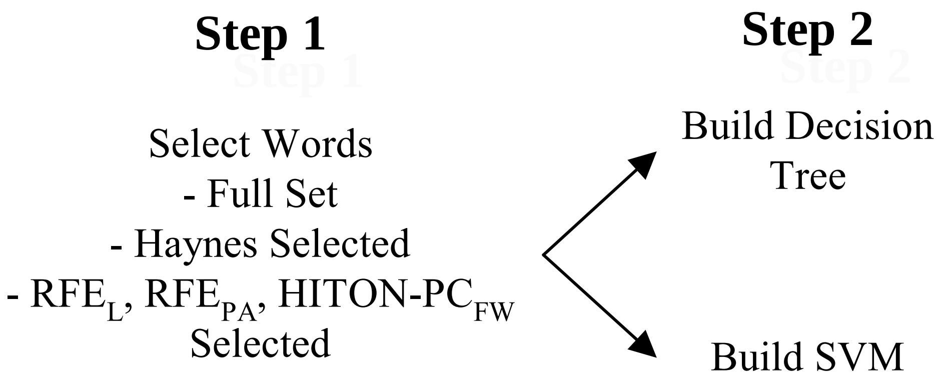
In order to significantly speed-up the algorithm we modify HITON in two ways: (a) we apply, as a first step, a univariate association-based reduction in the number of terms used (which was shown in [25] to lead to excellent classifiers - but not optimally small ones) and (b) we do not pursue full induction of the Markov Blanket (i.e., parents, children and spouses of the Target category in the Bayesian Network representing the classification tasks) but use an approximation to the Markov Blanket the parents and children only.

The price paid for the resulting speed up is that the modified algorithm is no longer sound even if the original HITON assumptions hold. This is because some members of the Markov Blanket (i.e., parents of children that do not have direct arcs with the target variable) will be omitted; yet they are necessary for optimal classification in the worst case. As we will see this heuristic modification to HITON works well in our experiments.

Experimental Design

The design is a simple 2 step methodology. In step 1, a word set is selected to represent the domain. In step 2, an SVM classifier and a decision tree classifier are trained using this word set.   This design is illustrated in Figure II-4.

Figure II-4 – Experimental Design Methodology



For step 1, we used 3 sets of words as inputs to the decision tree: the word set with the best performance/ feature ratio from each of the 3 selection methods, the full word set, and the word set from the Haynes experts [8].

The decision trees and the subsequent Boolean queries are evaluated quantitatively via a combined sensitivity-specificity measure to the CQF filters of Pubmed; they are also examined qualitatively.

Results

The performance to feature results are shown in Table II-5. We use SVMs and examine the area under the receiver operating curve (AUC). The Markov blanket HITON-PC_{FW} algorithm has the best performance-to-feature ratio and is able to reduce from 27891 features to 13.

Table II-5 – Feature Selection Performance

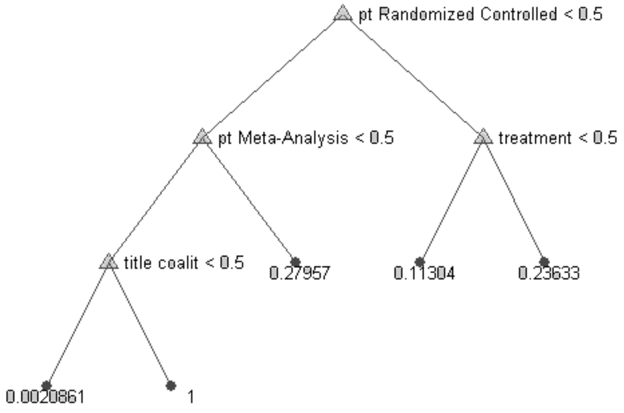
HITON-PC$_{FW}$ (13 Features)*				0.92 AUC		
RFE						
Features	**28000**	**1743**	**871**	**217**	**54**	**13**
RFE$_L$	0.95	0.85	0.96	0.97	0.86	#
RFE$_{PA}$	0.83	0.95	0.94	0.95	0.92	0.91

* - HITON-PC$_{FW\_}$ returns a single set.

# - RFE$_L$ did not converge to a solution

The performance of decision trees with varying inputs are shown in Table II-6. The best decision tree (best AUC performance) is illustrated in Figure II-5.

Figure II-5 – Decision tree produced by HITON/ DT



The triangles are decision nodes. The left branch corresponds to the word being absent, and the right branch to the word being present. The leaves indicate the probability of a high quality treatment related document.

Table II-6 – Decision Tree Performance on Test Set

Method	AUC	Words in pruned tree
Full Feature Set (27891 features)		
- SVMs	0.98	N/A
- DT	0.94	2
HITON-PC$_{FW}$ Feature Set (13 features)		
- SVM	0.95	N/A
- DT	0.95	4
Haynes Feature Set (747 features)		
- SVM	0.94	N/A
- DT	0.93	2

Table II-6 shows that the best performing decision tree is using the HITON-PC$_{FW}$ feature set. The other decision tree methods follow closely. The words in the trees differ. Using the full feature set, the terms "publication type (pt) randomized controlled trial (RCT)"(top node) and "pt meta-analysis" are returned. Using the Haynes feature set, the terms "pt RCT" (top node) and "mesh heading RCT" are returned. The terms using the HITON-PC$_{FW}$ feature set are in Figure II-5.

Table II-7 compares the CQF filters with the decision trees. For each constructed decision tree we measure the sensitivity and specificity for the given task. These statistics provide two related measures for comparing two algorithms. None is sufficient by itself. This is because an algorithm may achieve perfect sensitivity by classifying all samples as positive or perfect specificity by classifying all samples as negative. Thus, a combined measure is required. The measure we used is the proximity of the sensitivity and specificity of the algorithm to perfect sensitivity and specificity expressed as [26]:

$$dist = \sqrt{(1 - sens)^2 + (1 - spec)^2}$$

Table II-7 – Decision Trees Compared to CQF filters

Method	**Distance**
CQF filter – optimized for sensitivity	0.23
CQF filter – optimized for specificity	0.50
Full feature set/ decision tree	**0.11**
HITON features set/ decision tree	**0.11**
Haynes feature set/ decision tree	**0.11**

Note, that we cannot use AUCs or fix the measures. First, AUC's cannot be generated for the CQF filters because the documents are not ranked. Either the query is satisfied or not. Second, fixing sensitivity and specificity as used in [9] cannot be used because of the limited thresholds output by the decision trees. Equivalent matches cannot be generated.

The decision tree methods (bolded) outperform both optimized CQF filters and have the best tradeoff between sensitivity and specificity.

In additional experiments, we ran the Flake method on this dataset. We found that the classifier performance was poor and selected counter-intuitive terms. Since the Flake method is highly heuristic and not designed for this domain, we did not pursue it further.

Discussion

Every decision tree method produces a tree that is manageable, readable, and can be validated by humans. The simplicity of the solutions is not surprising since, for this proof-of-concept study, we purposely chose a task that had simple guidelines.

Specifically, the decision tree in Figure II-5 has words that are intuitive to the treatment class. Publication type (pt_) randomized controlled trial and pt meta analysis seem appropriate considering the criteria of the ACP journal [13]. The ACPJ criteria for treatment are a random allocation of participants to comparison groups, 80% follow-up of those entering the study, and the outcome to be of known or probable clinical importance.

The Boolean queries at each leaf appear to be equally sensible. For example, the leaf obtained with pt randomized controlled trial with the word treatment in the abstract has a 24% probability of being a good document. Human experts could develop this Boolean query intuitively.

The next highest leaf is the article is *not* a pt randomized controlled trial, but is a pt meta analysis, then we are 28% sure that the article is of high quality in the treatment class. This query is less intuitive since it says that meta-analysis qualify as high quality treatment related articles.

In light of these Boolean queries, how easy would it be for an expert to construct them? The first query seems straightforward and is an example of a disjunctive query that experts excel at constructing. It follows closely the intuitive notion of what content bearing words would indicate a high quality treatment related study. We argue that second query is more difficult for an expert to construct. Experts can readily explain, in specific instances, what would make a good document, but when it comes to generating an efficient query, especially with “not” qualifiers, the problem of selecting the appropriate words becomes very hard [27, 28].

Given the good performance of the decision trees, it makes sense to ask why the feature selection process is necessary if running a decision tree using the full feature set produces good results? The answer is not apparent in this data set. Initial experiments in more complex categories showed that feature selection is necessary. We ran the same experimental design in the etiology category area, and preliminary results show that a decision tree approach on the full feature set does not produce as good results as the feature selection/ decision tree method.

Table II-8 motivates the use of the feature set/ decision tree method. In this category, the full feature built decision tree is less complicated, but does not perform as well. Using the HITON/ decision tree method, the tree is more complicated, but the tree performs better than the former method. Feature selection is, indeed, necessary in this more complex category.

Table II-8 – Etiology Decision Tree Comparison

Decision Tree Elements	**HITON/ DT**	**Full/ DT**
Number Features Used	13	85662
Max Depth	12	3
Number Leaves	45	5
Number Nodes	45	5
AUC performance	0.90	0.80

The methodology in the present paper reduces labor by learning the needed words from a corpus rather than asking experts to define words that represent the treatment area. This approach has two advantages. It reduces variability in term selection and bypasses the need for appropriate experts. While the opposite method of [8] produces results comparable to the corpus-based methods (Table II-6), the resulting tree is limited in its interpretation. For example, the best Boolean query is "not pt randomized controlled trial and mesh heading randomized controlled trial." This query essentially represents the same concept, and, in comparison to the second query, misses the pt meta analysis concept. The Boolean query construction of Haynes is sub-optimal for this category as seen by the distance measures in Table II-7, and lower AUC in Table II-6. The possibility of missing words that describe the content is a weakness in the methodology. Similarly human cognitive biases equating co-occurrence with association hinder the construction of effective Boolean queries by experts [28].

Conclusions

The contribution of this paper is 4-fold. First we have presented a combined feature selection/decision tree method that can produce decision trees that perform as well as the best text classifiers and outperform methods currently available for this task. Second, these decision trees are understandable, manageable, and amenable to validation by humans. Third, these trees and queries are generated automatically from a corpus hence the process can be readily repeated many times in similar domains/tasks. Fourth, the Boolean queries discovered can be readily applied in existing search engines.

Our future research will also explore this method in more difficult categories with broader criteria for ACP inclusion such as diagnosis, prognosis, and etiology to further delineate the limits of the methodology presented here as well as potential improvements.

Acknowledgments

Yin Aphinyanaphongs is supported from NLM training grant: T15 LM07450-01. The authors would like to thank Ioannis Tsamardinos and Alexander Statnikov for help with the feature selection methods and useful discussions.

References

1. Davidoff, F., et al., *Evidence Based Medicine: A New Journal To Help Doctors Identify the Information They Need.* BMJ, 1995. **310**: p. 1085-6.

2. www.mdconsult.com

3. www.up2date.com

4. http://www.ncbi.nlm.nih.gov/PubMed/

5. Shojania, K.G. and L.A. Bero, *Taking Advantage of the Explosion of Systematic Reviews: An Efficient MEDLINE Search Strategy.* Effec Clin Prac, 2001. **4**(4): p. 157-159.

6. Robinson, K.A. and K. Dickersin, *Development of highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed.* Int Epi Assoc, 2002. **31**: p. 150-153.

7. Bachmann, L., et al., *Identifying Diagnostic Studies in MEDLINE.* JAMIA, 2002. **9**(6): p. 653-658.

8. Haynes, B., et al., *Developing Optimal Search Strategies for Detecting Sound Clinical Studies in MEDLINE.* JAMIA, 1994. **1**(6): p. 447-458.

9. Aphinyanaphongs, Y. and C.F. Aliferis. *Text Categorization Models for Retreival of High Quality Articles in Internal Medicine*. in *AMIA*. 2003. Wash, D.C.

10. Silverstein, C., et al. *Analysis of a Very Large Web Search Engine Query Log*. in *SIGIR FORUM*. 1999.

11. Flake, G., et al. *Extracting Query Modifications from Nonlinear SVMs*. in *Int WWW Conf*. 2002. Honolulu, HA.

12. Schapire, R.E. *Theoretical views of boosting and applications*. in *Tenth International Conference on Algorithmic Learning Theory*. 1999.

13. *Purpose and Proc.* ACP Journal, 1999. **131**(1): p. A15.

14. Jerome, R.N., et al., *Info Needs of clinical teams: analysis of questions received by the Clinical Informatics Consult Service.* Bull Med Libr Assoc, 2001. **89**(2): p. 177-184.

15. Burges, C., *A tutorial on support vector machines for pattern recognition.* Data Mining and Knowledge Discovery, 1998. **2**: p. 121-167.

16. www.mathworks.com

17. http://www.cis.tugraz.at/igi/aschwaig/software.html

18. Joachims, T., ed. *Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning.*, ed. B. Scholkopf, C. Burges, and A. Smola. 1999, MIT-Press.

19. Osuna, E., R. Freund, and F. Girosi. *Training support vector machines: an application to face detection*. in *Conf on Computer Vision and Pattern Recognition*. 1997.

20. Apte, and Weiss, *Data Mining with Decision Trees and Decision Rules.* Future Gener Computer Systems, 1997.

21. Murthy, S., *Automatic Construction of decision trees from data: A multi-disciplinary survey.* Data Mining and Knowledge Discovery, 1997.

22. Duda, R., P. Hart, and D. Stork, *Pattern Classification*. 2nd ed. ed, ed. J.W. Sons. 2001.

23. Guyon, I., et al., *Gene Selection for Cancer Classification using Support Vector Machines.* Machine Learning, 2002. **46**: p. 389-422.

24. Aliferis, C.F., I. Tsamardinos, and A. Statnikov. *HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection*. in *AMIA*. 2003. Washington, DC.

25. Yang, Y. and J. Pederson. *A comparative study on feature selection in text categorization*. in *14th International Conference on Machine Learning*. 1997: M. Kauffman.

26. Tsamardinos, I., C.F. Aliferis, and A. Statnikov. *Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations*. in *Proceedings of the 9th ACM SIGKDD International Conference on KDD*. 2003.

27. Bourne, L.E., Jr. and D.E. Guy, *Learning Conceptual Rules: II The role of positive and negative instances.* Journal of Experimental Psychology, 1968. **77**: p. 488-494.

28. Plous, S., *The Psychology of Judgement and Decision Making*. 1993, McGraw-Hill Inc: New York.

CHAPTER III

III. MODELS AND EVALUATION OF RETRIEVAL PERFORMANCE

In this section, I explore the machine learning filter models when compared to other methods to identify high quality articles in the literature. Specifically, I compare the machine learning models built for the specific gold standard to bibliometric citation count, impact factor, and non-specific machine learning models to rank the literature. In subsequent research, I compare the machine learning models to total web page hit count for each article, 2005 impact factors, bibliometric citation count, Google Pagerank, and Yahoo Webranks for ranking the literature.

A Comparison of Citation Metrics to Machine Learning Filters for the Identification of High Quality MEDLINE Documents

Aphinyanaphongs Y, Statnikov A, Aliferis C. "A Comparison of Citation Metrics to Machine Learning Filters for the Identification of High Quality MEDLINE Documents." J American Medical Informatics Association. 2006; 13 (4): 446- 455.

Abstract

OBJECTIVE: The present study explores the discriminatory performance of existing and novel gold-standard-specific machine learning (GSS-ML) focused filter models (i.e., models built *specifically* for a retrieval task and a gold standard against which they are evaluated) and compares their performance to citation count and impact factors, and non-specific machine learning (NS-ML) models (i.e., models built *for a different task and/or different* gold standard).

DESIGN: Three gold standard corpora were constructed using the SSOAB bibliography, the ACPJ-cited treatment articles, and the ACPJ-cited etiology articles. Citation counts and impact factors were obtained for each article. Support vector machine models were used to classify the articles using combinations of content, impact factors, and citation counts as predictors.

MEASUREMENTS: Discriminatory performance was estimated using the area under the receiver operating characteristic curve and n-fold cross-validation.

RESULTS: For all three gold standards and tasks, GSS-ML filters outperformed citation count, impact factors, and NS-ML filters. Combinations of content with impact factor or citation count produced no or negligible improvements to the GSS machine learning filters.

CONCLUSIONS: These experiments provide evidence that when building information retrieval filters focused on a retrieval task and corresponding gold standard, the filter models have to be built specifically for this task and gold standard. Under those conditions, machine learning filters outperform standard citation metrics. Furthermore, citation counts and impact factors add marginal value to discriminatory performance. Previous research that claimed better performance of citation metrics than machine learning in one of the corpora examined here is attributed to using machine learning filters built for a different gold standard and task.

Index Terms: Information Retrieval, Pubmed, Machine Learning, Artificial Intelligence.

Introduction & Background

The growth of publication volume in the majority of fields of biomedicine is rapidly becoming intractable. Modern approaches to biomedical information retrieval are seeking to alleviate the problem by developing specialized filters that find documents that satisfy special content or methodological criteria. Such filters have been developed, for example, to identify randomized controlled trials or to select documents that focus on prognosis and satisfy rigorous criteria of statistical design and analysis, etc. This Focused Filter Paradigm is implemented either via automated methods based on machine learning [1] or on manual and semi-manual construction of search queries tailored to the criteria of interest [2-4].

Citation metrics such as citation count and impact factor have a rich history in medical bibliometrics as indicators of impact, and indirectly of quality, of scientific papers [5, 6]. The recent successful application of advanced citation-based algorithms such as PageRank [7] and Kleinberg's HITS algorithms [8] in WWW search has re-invigorated interest in citation metrics for biomedical bibliographies.

Citation metrics differ dramatically from the focused filter paradigm in identifying documents. Citation metrics capture a document's research impact directly, and they may also serve as proxies for methodological quality or utility. Focused filters, in contrast, can, in principle, capture arbitrarily specialized and complex sets of quality criteria used by human editors to create a set of indexed documents. Because every focused filter is built for specific criteria (e.g., whether a document describes a randomized controlled trial or not), we would expect, a priori, focused filter models to outperform, with respect to the same criteria, a generic metric such as citation that is not devised for these criteria. For example, a paper describing a randomized controlled trial may have equal number of citations with a case-control study, rendering the two non-distinguishable by citation number whereas from a pattern recognition or from a biomedical librarian's perspective the citations are perfectly distinguishable.

Very recent research by Bernstam et al, used a bibliographic collection of articles selected by the Surgical Oncological Society for their "importance" in surgical oncology and reported - counter to the above intuitive principle - that citation count ranks documents from this surgical oncological bibliography (SSOAB gold standard) higher than documents ranked with PageRank or by a machine learning (focused filter) model [9]. However, [9] evaluated machine learning models that were *not* built specifically either for the quality criteria of SSOAB or for its content type, but rather for *different* quality and content criteria (evidence based medicine quality criteria captured by the ACP Journal Club corpus gold standard [10]). In other words, focused filters with a different focus than the gold-standard and content type at hand were compared to citation metrics. The research in [9] therefore supports the claim that citation count, which is an easy to compute, context-free, and relatively accessible metric, may, in fact, be better for finding high-quality documents than sophisticated human or pattern recognition queries and models. A natural question to ask is whether the conclusions of [9] can be attributed to use of filters with a different focus, or whether, intrinsically, citation metrics are superior to machine learning filters (regardless of focus). Answering this question has great methodological importance since it will help indicate, in part, what approaches are likely to yield better results in developing next-generation biomedical information retrieval systems.

Hypothesis & Experiments

Our main hypothesis is that citation metrics are not superior to focused filter models as long as the latter are built for the specific criteria used to evaluate them. We conducted a series of experiments to test this hypothesis:

- **Experiment 1**: We built content-based, (i.e., title, abstract terms, journal, MeSH terms) SSOAB-specific filter models using machine learning and compared them to citation-based models and content-based focused filters specific to the ACP Journal Club gold standard [10] using the SSOAB as the gold standard.

  In addition, we applied feature selection and an SVM-based feature weighting method to examine the implicit criteria used by the SSOAB editors in building their corpus.
- **Experiment 2**: We built machine learning models that, in addition to the document content, include citation metrics as predictors. We analyzed whether citation metrics add any value to classification of SSOAB documents compared to classification based on only content data.
- **Experiment 3:** We tested whether the performance of the machine learning filters is partially attributable to their predicting citation count. We specifically tested how well the machine learning modeling techniques used for the filters predict citation counts from the document content.

- **Experiment 4**: To further establish the generalizability of these results with other corpora and datasets, we repeated the above three experiment sets with two ACP Journal Club corpora in the treatment and etiology categories.

Methods

In section A, we specify the definitions used throughout the paper. In section B, C, and D, we explain the methods used to create the SSOAB and ACP Journal Club gold standards and obtaining their respective citation counts and impact factors. In Section E, we explain how the articles are represented and classified by the SVM classification model described in Section F for experiments 1, 2, and 4. In Section G, we describe the regression models used to predict citation count for experiment 3. In sections H and I, we describe the performance metrics and the cross-validation method used for performance estimation and model selection. Finally, in section J, we describe the feature selection methods used to understand the implicit criteria of the selected articles in the SSOAB gold standard.

Definitions

We introduce here definitions that are important for following the design, methods, results and conclusions of the paper. Throughout the paper, we use filter, models, and filter models interchangeably.

Definition 1. *Content-based filter*: A filter (human query or machine learning model) that is based on the content of the MEDLINE document. In the present study, the content includes the title, abstract, journal title, MeSH terms or combinations of them, represented by schemes appropriate to the modeling methodology.

Definition 2. *Context-free citation metric*: Any citation metric that is calculated independent of clinical or research context of use, or of gold standards of quality, importance, utility, cost, etc. Citation count, PageRank, and Impact Factor are context-free citation metrics.

Definition 3. *Gold-standard-specific (GSS) filter*: Any filter designed for, and evaluated by, a specific gold standard and/or related context of use. For example, a filter designed to identify rigorous treatment articles in internal medicine according to the ACPJ treatment methodological quality criteria.

Definition 4. *Non-specific (NS) filter*: Any filter designed for a specific gold standard and/or related context of use but used for a different context of use and/or evaluated by a different gold standard. For example, a filter designed to identify rigorous treatment articles in internal medicine according to the ACPJ treatment methodological quality criteria but used to find articles included in the SSOAB bibliography.

Gold Standard Construction

1. SSOAB

The SSOAB bibliography is a collection of articles selected by the Surgical Oncological Society for their “importance” in surgical oncology [11]. The bibliography includes 458 articles covering a wide range of topics and study designs in surgical oncology. The bibliography does not purport to be evidence-based in allowing only articles with high methodological rigor nor does it have strict inclusion criteria by the editors. We emphasize that in light of the lack of stated editorial standards of the SSOAB corpus, we are interested in it primarily because this corpus serves as the basis for the methodological evaluations and resulting claims in [9] which is central to the main hypothesis of the present study.

The SSOAB corpus was constructed as follows: we began with the 458 articles as positives and augmented the corpus with negative articles. We identified negative documents by examining the journal and issue for each published article included in the SSOAB bibliography, and taking all *other* original research articles not selected by the SSOAB with abstracts (as indexed by Pubmed) *in the same journal and issue* to be negative instances. This procedure generated a corpus that consists of "pure positive" documents (i.e., ones included in the SSOAB) and "pure negative" documents (i.e., following the rationale that documents we characterized as negative using this process cannot be falsely negative since at least one positive article was identified as positive in the same issue, the remaining articles were assumed reviewed and are truly negative and not negative by omission) [1]. We further excluded 27 of the original 458 positive articles that did not have available abstracts from Pubmed. These methods resulted in an SSOAB corpus with 431 positives and 7,379 negatives.

2. ACPJ-treatment, ACPJ-etiology

[1] In [9], it is proposed that only the documents in the SSOAB are the true positives and all else are negatives. This is a non-sequitur in the context of that study's conclusions since one would only need a look-up table to find the good articles, and not citation (or other) metrics as recommended by [9]. In other words, since all the documents are assumed labeled by [9], the use of citation or any predictive method for identifying articles is unnecessary. An ideal design would be to rank the articles by citation count and observe how citation predicts *new* SSOAB editor inclusion/exclusion decisions. Implicit thus in [9] is that the SSOAB positives are a subset of all good articles and that "SSOAB-positive-like" documents will be returned when using citation count as filtering criterion.

The ACP Journal Club is a highly-rated meta-publication. Every month expert clinicians review a broad set of journals in internal medicine, and select articles in these journals according to specific criteria in the content areas of treatment, diagnosis, etiology, prognosis, quality improvement, clinical prediction guide, and economics. Selected articles are further subdivided into articles that are summarized and abstracted by the ACP because of their “clinical importance”, and those that are only cited because they meet all the quality selection criteria but may not pertain to vitally “important clinical areas”. For the purposes of the present study, articles were abstracted or cited by the ACP are considered positive instances and all other articles in the same journals were considered negative. The criteria for inclusion in ACPJ can be found in [10].

We used for the present study a modified version of the ACPJ corpus as in [1]. We considered all articles cited and abstracted in the treatment and etiology categories from 49 selected journals covered by the ACPJ between July 1998 and August 1999 as positives, and all other articles published in the same 49 journals in the same period but not cited or abstracted as negatives. This procedure resulted in 15,786 documents with 205/15,581 positives/negatives in etiology and 379/15,407 in treatment respectively. Note that the method to build the ACPJ corpus differs from the SSOAB method in that the ACPJ documents are not limited to a specific issue, but instead to the documents published in a given time frame for the specific journal.

A. Citation Count

Citation count is the number of publications citing an article. We downloaded citation counts from the Web of Science [12] using a screen scraping interface coded in Python. The screen scraper established an http connection to the Web of Science servers and navigated through several GET and POST requests to identify an article and parse out the number of cited articles. We obtained citation counts of articles in the SSOAB [11] and ACPJ gold standards [1, 10] on August 2005 and August 2002. These collection dates allowed approximately 2.5-3 years for citations to accumulate in each respective gold standard.

A relatively small number of articles did not have a citation count since the corresponding journals were not followed by the Web of Science. For the SSOAB gold standard, we obtained 7,676 citations with counts out of 7,810 citations in the gold standard. For the ACPJ gold standard, we obtained 13,279 citations that had counts out of 15,786 citations in the gold standard.

For the articles without citation counts, we used the following imputation procedure to provide an estimate for the missing citation count values. For each article X with a missing citation count, we randomly selected an article Y with an observed citation count from the same labeled class and assigned the citation count of Y to X. We did not assign the mean citation count of each respective class as the citation count for articles with missing citations, because the machine learning algorithm would inappropriately use the assigned mean citation count as a near-perfect, but biased, predictor for classification of all documents with missing values.

B. Impact Factor

An impact factor of a journal is the average number of citations an article published in this journal receives in two years [13]. For example, the 2004 impact factor for journal X would be the number of citations received by articles published in X within 2002-03, divided by the total number of published articles in X within 2002-03. We obtained impact factors from the Web of Science for 2005 and 2001. These years corresponded to the time periods covered by the gold standard corpora.

C. Document Representation and Pre-processing for Machine Learning

The conversion of documents to a format suitable for the machine learning algorithms followed the procedures in [1]. The articles in the SSOAB and ACPJ selected journals were cross-referenced in PubMed, and the title, abstract, journal, and MeSH terms were extracted. We represented each document as a set of terms for the learning algorithms [14]. We additionally stemmed each term [15], removed "stopword" terms [16], and removed any terms occurring in fewer than 5 documents. Very infrequent terms are difficult to assess statistically and may affect negatively the generalization of the classification models. Selected terms from the title, abstract, and MeSH were further encoded as weighted features using a log frequency with redundancy scheme for all documents [17]. The SSOAB collection contained, after imputation of citation counts, 7,810 articles with abstracts and citation counts represented by 16,441 features including citation count. The ACPJ etiology and treatment collection contained, after imputation of citation counts, 15,786 articles with abstracts and citation counts represented by 28,229 features including citation count (see Gold Standard Construction section for additional information).

D. Classification Methods

In our experiments, we employed Support Vector Machine (SVM) classification algorithms. The SVM's calculate a maximal margin hyperplane separating two or more classes of the data. To accomplish this, the data are mapped to a higher dimensional space by means of a kernel function, where a separating hyperplane is found by solving a constrained quadratic optimization problem [18]. We used SVMs, because for several published text categorization tasks, SVMs have had superior classification performance compared to other methods [1, 19], and this motivated our use of them. We used an SVM classifier implemented in libSVM v2.8 [20] with a polynomial kernel. We optimized the SVM penalty parameter C over the range {0.1, 1, 10, 100} and degree d of the polynomial kernel over the range {1, 2, 3, 4}. Since theoretical literature on domain characteristics as it relates to optimal parameter selection is not yet developed , the ranges of costs and degrees for optimization were chosen based on previous empirical studies [1, 19, 21]. Different combinations of costs and degrees were exhaustively evaluated by cross-validation, and the best performing model was selected for the final application of the SVM classifier (see section on Performance Estimation and Model Selection).

E. Regression Methods

In our experiments for the citation prediction task, we used epsilon - Support Vector Regression (e-SVR) [22]. This regression technique uses an epsilon-insensitive loss function (as opposed to a square loss function in linear regression) to calculate an optimal surface that approximates the continuous response variable. Similar to SVM for classification, the data is mapped to a higher dimensional feature space by means of a kernel function, and the optimal approximating surface is found by solving a constrained quadratic optimization problem. We used e-SVR with a polynomial kernel implemented in libSVM v.2.8 [20]. We optimized the e-SVR penalty parameter C over {50, 100}, the kernel degree d over {1, 2, 3}, and used the software default epsilon of 0.1.

F. Performance Metrics

Among the many classifier performance metrics such as precision, recall, average 11-point precision, F[1] score, breakeven point, accuracy, error, and area under ROC curve (AUC) that have been used for two-class text categorization (for example, see [21, 23-25]), we decided to use AUC [26, 27] for the following two reasons. First, the AUC metric does not correspond to a single threshold on the classifier predictions which is the case for precision, recall, accuracy, F[1] score, and other common metrics (but not for average 11-point precision). The AUC is a comprehensive metric and is computed for values of sensitivity and specificity over all possible thresholds observed in the data. Second, unlike all other performance metrics mentioned above, AUC is insensitive to the class distribution [26]. Thus, the interpretation of AUC is fairly straightforward for this task[2]. Relying on performance measures that are sensitive to class distributions may be a misleading measure of discriminatory performance. For example, we would not use accuracy (defined as the proportion of correct classifications over all classifications) as a performance measure, because excellent accuracy can be achieved in extremely skewed distributions by classifying all documents as belonging to the most prevalent class [28].

In order to generate an ROC curve for classification experiments, we used outputs of the SVM model corresponding to distances from the testing examples to the maximum margin hyperplane that separates positive and negative training examples. The SVM outputs were ranked, and an ROC curve was generated from this ranked list of examples. The ROC curve for citation count was similarly determined by ranking the articles by citation count. The area under ROC curve was computed as in [27].

[2] AUC changes between 0 and 1 with 1 being perfect classification, 0.5 being random classification performance, and 0 being inverse classification with all true positives classified as negatives and all true negatives classified as positives [26].

For the experiments with regression algorithms, we used Pearson's and Spearman's correlation coefficients [29] to measure how well the predicted citation count matches the true citation count. We also used R^2 (also known as "coefficient of determination") which indicated the proportion of variance in the true citation count accounted for by the regression model [29]. In the statistical literature, correlation coefficients greater than 0.8 (i.e., R^2> 0.64) are generally considered as indicative of strong correlation, whereas a correlation smaller than 0.5 (i.e., R^2<0.25) is generally considered as weak.

G. Performance Estimation and Model Selection

We used 5-fold cross-validation to estimate the performance of the learning algorithms [30]. This procedure first divided the data randomly into 5 non-overlapping subsets of documents where the proportion of positive and negative documents in the full dataset is preserved for each subset. Next, the following was repeated 5 times: we used one subset of documents for testing (the "original testing set") and the remaining four subsets for training (the "original training set") of the classifier. The average performance over 5 original testing sets is reported.

In order to optimize parameters of the SVM or epsilon-SVR algorithms, we used another "nested" loop of cross-validation by further splitting each of the 5 original training sets into smaller training sets and validation sets. For each combination of learner parameters, we obtained cross-validation performance and selected the best performing parameters inside this inner loop of cross-validation. We next built a model with the best parameters on the original training set and applied this model to the original testing set. Details about the "nested cross-validation" procedure can be found in [31, 32]. Notice that the final performance estimate obtained by this procedure will be unbiased because each original testing set is used only once to estimate performance of a single model that was build by using training data exclusively.

H. Feature selection and feature weighting for examining implicit criteria used in gold standard corpus.

The SSOAB corpus was not built using a set of explicit criteria like the ACPJ corpus [10]. To gain insight into the implicit criteria used for the SSOAB, we performed feature selection and ranked the selected features according to their contribution weight to an SVM classification model built with only these features for predicting class membership (i.e., SSOAB inclusion or not).

In general, there exist many feature selection algorithms applicable to text categorization. We focus here on Markov Blanket induction ones such as HITON [33] because under the broad distributional assumption of Faithfulness, they find a unique and smallest set of predictors that gives the largest predictive performance for “universal approximator” learners such as SVMs [34]. To speed up the feature selection operation, we used the HITON_PC algorithm which approximates the Markov blanket.

Specifically, while the Markov Blanket (the provably minimal set of optimal predictors) consists of the set of parents, children, and spouse nodes of the response variable in the Bayesian network that is a perfect map of the dependencies and independencies in the joint probability distribution of predictor terms and the response variable (target class), HITON_PC is guaranteed to return the parents and children of the target variable and has been shown in prior experiments to approximate well the Markov Blanket in text categorization tasks while being more computationally efficient than finding the latter [35]. We used an implementation of HITON_PC from the *Causal_Explorer* toolkit [36] with G^2 statistical test and a threshold p-value of 0.10.

HITON_PC was executed on binary features indicating presence or absence of a term in the document

The entire procedure for understanding the importance of terms for inclusion in the SSOAB has the following three steps:

I. Features were selected by HITON_PC in the context of cross-validation design for each original training set. Using the data corresponding only to selected features, the SVM classifier was optimized and trained on the original training set and tested on the original testing set (see Performance Estimation and Model Selection section). This allowed us to access classification performance of selected features in an unbiased fashion since the testing data is neither used for classifier learning nor for feature selection.

II. If the performance of HITON_PC features matched one of the entire feature set (i.e. without feature selection, which is best case), then we (a) re-selected features using all examples in the corpus and (b) optimized and trained the SVM classifier on the selected features using all examples in the dataset. Notice that we can use all data in this analysis since the analysis is explanatory and not predictive.

III. Finally, we computed contribution Δ_i of each selected feature *i* on step II to the SVM model's objective function as described in [37]. We report the normalized contribution of each feature which is equal to $\Delta_i / \sum\Delta_i$.

Results

Experiment 1: The results for experiment 1 are shown in Table III-1. GSS focused filter models built using machine learning for the SSOAB gold standard out-performed impact factor, citation counts, and NS models with a different focus in predicting SSOAB article inclusion. The GSS focused filter models built specifically for content have the highest AUC of 0.893. Prediction by impact factor in both 2001 and 2005 were nearly random at 0.549 and 0.558 AUC, respectively. Citation count by itself was moderately predictive with an AUC of 0.791. Predictions using NS models built for the ACP Journal Club treatment category were nearly random at 0.548 AUC.

Table III-1 – Comparison of gold-standard-specific, content-based machine learning filters with citation metrics and models built for ACPJ criteria in the SSOAB quality classification task

Gold standard: SSOAB	**Area under the ROC curve**	**p – value***
SSOAB-specific (GSS) filters	0.893 (weighted)	N/A
Citation Count	0.791 (ranked)	<0.0001
ACPJ Treatment-specific (NS) filters	0.548 (weighted)	<0.0001
Impact Factor (2001)	0.549 (ranked)	<0.0001
Impact Factor (2005)	0.558 (ranked)	<0.0001

weighted – content terms weighted by log frequency with redundancy scheme [17].

ranked – citations are ranked by counts (or impact factor) and a composite ROC generated.

* - p-values for each feature set are calculated in comparison to the content only focused filters using the Delong paired comparison test [38].

The AUC produced by the content method alone were significantly different than the AUC produced by the citation metrics (using the Delong AUC paired comparison statistical test at the 0.05 level [38]). Results indicate that using machine learning models built for a specific gold standard is essential for discriminative performance in this task. The SSOAB specific models outperformed the ACPJ specific models by 0.345 when applied to the SSOAB corpus.

Additional Analyses: Feature Selection and term importance

We performed feature selection and feature weighting experiments to gain insight into the SSOAB corpus construction. The results of feature selection and weighting are presented in Table III-2.

Table III-2 – Features selected by the HITON_PC algorithm from the entire SSOAB corpus (i.e. using all documents). Some words are stemmed [15].

Feature Rank	Features	Normalized contribution
1	adjuv	0.222
2	Pancreatic Neoplasms[MeSH]	0.18
3	node	0.134
4	cutan	0.115
5	randomis[Title]	0.078
6	Minnesota[MeSH]	0.043
7	pancreaticoduodenectomi	0.034
8	discov[Title]	0.03
9	N Engl J Med[Journal]	0.029
10	resect	0.026
11	referr	0.02
12	cancer	0.017
13	melanoma[Title]	0.016
14	soft[Title]	0.016
15	carcinoma	0.014
16	surgery[MeSH]	0.01
17	North America[MeSH]	0.005
18	Stomach:pathology[MeSH]	0.003
19	Multiple Endocrine Neoplasia:genetics[MeSH]	0.003
20	Hospitals, Veterans[MeSH]	0.002
21	Animals[MeSH]	0.001
22	Metabolism[MeSH]	0.001
	Performance of SVM with the above 23 features	0.834
	Performance of SVM with all features (16440 features)	0.893

These results are interesting since the SSOAB editors are not operating with explicit selection criteria. The selected words are indicative of the unstated criteria and may reveal possible biases (positive and negative ones) in article selection by the SSOAB. The top 5 words suggest that the SSOAB editors were selecting articles that are related to surgical oncology, are treatment related (through the inclusion of “randomized”), and are biased toward pancreatic neoplasms. Inspection of words ranked 6-16 further support the selection of surgical oncological articles with a bias to articles discussing pancreaticoduodenal cancer, articles with studies taking place in Minnesota, and articles published in the New England Journal of Medicine, while the 6 lowest weighted words account for less than 0.015 of the classifier’s behavior and their interpretation is not as important.

We extended the analysis by inspecting “stable” features by taking the intersection of the selected words from each cross-validation training set. This procedure resulted in 8 most stable features (“resect”, “node”, “surgery[MeSH]”, “adjuv”, “cancer”, “Pancreatic Neoplasms[MeSH]”, “randomis[Title]”, “N Engl J Med[Journal]”). These words further support our observations of article selection by the SSOAB with biases toward pancreatic neoplasms and publications by the New England Journal of Medicine.

This feature analysis is not exhaustive or conclusive, and the results are included to illustrate that techniques are available to analyze the corpora to detect terms significant for document selection in each corpus. For a previously published analysis of term importance for ACPJ, please see [35].

Experiment 2: The results of experiment 2 are shown in Table III-3. Machine learning GSS focused filters that include citation metrics as predictors are minimally better than filters that do not. The addition of citation information to the content models increased the AUC by 0.022 over using content alone. The resulting AUCs were statistically different when comparing content to content + citation count using the Delong method at the 0.05 level [38].

Table III-3 – Comparison of gold-standard-specific, content-based machine learning filters with hybrid content + citation metric models

Gold standard: SSOAB	**Area under the Curve**	**p – value***
SSOAB-specific model (from experiment 1)	0.893 (weighted)	N/A
SSOAB-specific model (GSS Content + Citation Count-based)	0.915 (weighted + normalized)	<0.0001
SSOAB-specific model (GSS Content + Impact Factor (2005) – based)	0.899 (weighted + normalized)	0.026
SSOAB-specific model (GSS Content + Citation Count + Impact Factor (2005) – based)	0.914 (weighted + normalized)	<0.0001

weighted – content terms weighted by log frequency with redundancy scheme [17].

normalized – citation counts and impact factors are normalized between 0 and 1 and added as a feature.

* - p-values for each feature set are calculated in comparison to the content only focused filters using the Delong paired comparison test [38].

Additionally, including impact factor with citation count and content showed no improvement in area under the curve when compared to using content with citation count (since impact factor is a composite measure of of citation count) . Content alone compared to content with impact factor showed a statistically significant, but negligible improvement in AUC.

Experiment 3: Table III-4 shows the results for experiment 3. Correlation coefficients of 0.46 (R^2 of 0.212) for the SSOAB citation prediction task showed limited ability for content to predict citation count. Similar results were provided for the ACP Journal Club gold standard in both etiology and treatment categories: the predictions had small correlations of 0.60 and 0.61 for Pearson (R^2 of 0.360 and 0.372 respectively) and 0.49 for Spearmans correlation coefficients. The inability of SVM models to predict well citation counts means that citation count contains information not captured by the machine learning model (by the results of experiment 3); however, the information captured by citation counts do not add to the classification that is based on content alone (as evidenced by the results of experiment 2).

Table III-4 – Results of Support Vector Regression prediction of citation counts from content compared to the true citation counts in all corpora.

Corpus	**Pearson Corr. Coef. of predictions with true citation count**	**Spearman Corr Coef. of predictions with true citation count**	**Coefficient of Determination (R^2)**
SSOAB	0.46	0.46	0.212
ACPJ Etiology	0.60	0.49	0.360
ACPJ Treatment	0.61	0.49	0.372

Experiment 4: Table III-5 and Table III-6 provide results analogous to experiments 1 and 2 but for the ACP Journal club categories in etiology and treatment. In etiology, with results shown in Table III-5, the GSS focused machine learning models outperformed the citation methods and the NS models built using the SSOAB corpus. Also, the inclusion of citation metrics with content did not add value relative to a strictly content-based model. The content based model achieved AUC of 0.932 outperforming citation count, impact factors, and the SSOAB-based NS model that have AUCs of 0.691, 0.670, 0.673, and 0.772 respectively (see Table III-5). The addition of citation count or impact factor with content models did not improve discriminatory performance noticeably.

Similar results are shown for the treatment task in Table III-6. The GSS content based models outperformed any individual citation method and the NS models built using the SSOAB corpus. The models including citation metrics with content did not add value. The GSS content based models gave an AUC of 0.966, and the inclusion of citation metrics did not add value to the classification. Citation is moderately predictive at AUC of 0.762 and impact factors for 2001 and 2005 are even less so with AUCs of 0.601 and 0.594 respectively. The SSOAB specific models applied to the ACPJ treatment category gave an AUC of 0.770. In both ACPJ categories, inclusion of impact factors as a predictor did not improve classification performance. The results of these experiments showed that the GSS focused filter's advantage over citation metrics and NS models built for other gold standards generalizes beyond the SSOAB corpus.

Study Limitations

The current work compares citation metrics with machine learning ones on the same gold standard (SSOAB) just as [9] does. Despite its limitations of not using explicit inclusion criteria and of not being updated very regularly, we included SSOAB primarily because it allows us to compare to the results and conclusions of [9], a comparison central to the main hypothesis of the

Table III-5 – Comparison of ACPJ_etiology-specific content-based machine learning filters with citation metrics in the ACPJ-Treatment quality classification task.

Gold standard: ACPJ Etiology	**Area under the Curve**	**p – value***
ACPJ_etiology-specific filter (GSS, Content-based)	0.932 (weighted)	N/A
Citation Count	0.691 (ranked)	<0.0001
Impact Factor (2001)	0.670 (ranked)	<0.0001
Impact Factor (2005)	0.673 (ranked)	<0.0001
ACPJ_etiology -specific filter (GSS Content + Citation Count –based)	0.935 (weighted + normalized)	0.05
ACPJ_etiology -specific filter (GSS Content + Impact Factor (2005) –based)	0.924 (weighted + normalized)	<0.0001
ACPJ_etiology -specific filter (GSS Content + Citation Count + Impact Factor (2005) –based)	0.928 (weighted + normalized)	0.04
SSOAB-specific Models (NS, Content-based Only)	0.772 (weighted)	<0.0001

weighted – content terms weighted by log frequency with redundancy scheme [17].

normalized – citation counts and impact factors are normalized between 0 and 1 and added as a feature.

ranked – citations are ranked by counts (or impact factor) and a composite ROC generated.

* - p-values for each feature set are calculated in comparison to the content only focused filters using the Delong paired comparison test [38].

Table III-6 - Comparison of ACPJ_treatment-specific, content based machine learning filters with citation metrics in the ACPJ-Treatment quality classification task

Gold standard: ACPJ Treatment	**Area under the Curve**	**p – value***
ACPJ_ treatment-specific filter (GSS, Content-based)	0.966 (weighted)	N/A
Citation Count	0.762 (ranked)	<0.0001
Impact Factor (2001)	0.601 (ranked)	<0.0001
Impact Factor (2005)	0.594 (ranked)	<0.0001
ACPJ_ treatment-specific filter (GSS Content + Citation Count –based)	0.966 (weighted + normalized)	0.15
ACPJ_ treatment-specific filter (GSS Content + Impact Factor (2005) –based)	0.962 (weighted + normalized)	<0.0001
ACPJ_ treatment-specific filter (GSS Content + Citation Count + Impact Factor (2005) –based)	0.963 (weighted + normalized)	<0.0001
SSOAB-specific Filters (NS, Content-based Only)	0.770 (weighted)	<0.0001

weighted – content terms weighted by log frequency with redundancy scheme [17].
normalized – citation counts and impact factors are normalized between 0 and 1 and added as a feature.
ranked – citations are ranked by counts (or impact factor) and a composite ROC generated.
* - p-values for each feature set are calculated in comparison to the content only focused filters using the Delong paired comparison test [38].

present paper. While we use the gold standard and metrics that [9] employ, our research design also differs in the following specific ways:

- First, in the interest of generality we test our hypotheses not only on one corpus (i.e., SSOAB) but also on the ACPJ-treatment and ACPJ-etiology gold standards. Hence our results have a greater degree of generality.
- Second, [9] apply the metrics and filters on all of Medline; whereas, we train models on a part of Medline and test the models on an independent subset. We believe that this difference corresponds to what we perceive as a non-trivial flaw in [9]: by testing performance on all of Medline, [9] does not allow for generalizing the performance of their metrics and models. In effect their design amounts to that, among all Medline documents, only a few hundred ones included in SSOAB are of "importance" in surgery. Further, solving this problem exactly is rather trivial: just maintain a lookup table with all SSOAB positive articles. However with the design of the present study we address the issue of generalization beyond the studied SSOAB documents: can we show that filtering mechanisms or criteria/metrics can identify "SSOAB positive-like" documents in the future? (rather than simply regurgitating the known SSOAB positive ones?). The current design that uses separate training and testing document collections allows us to answer this question.
- Third, [9] uses HITS and precision-recall curves for a limited set of queries. We are using area under the ROC curve (AUC). We preferred AUC because both HITS curves and precision-recall curves are affected by the prevalence of positive documents in the corpus especially as this prevalence is sensitive to the choice of query and a priori is expected to vary considerably from query to query. [9] uses 40 queries that are by no

means standard in the field and are not necessarily capturing properties of real-life queries. They normalize and average their results over the 40 queries. In contrast, in the present study, we compute AUC performance not for a specific query set but for all the corpus (which is to effectively be interpreted as an average over all possible queries). Our findings indicate that the findings by [9] hold in this more general design as well so they are not necessarily an artifact of their experimental design. We have conducted additional experiments that provide HITS and precision-recall results in our cross-validation design. As expected, the results are consistent with our ROC results, and we include them in the appendix. The graphs in the appendix should only be interpreted in the context of the experiments in this paper, and not be compared to the hits and precision-recall curves in [9] due to differences in priors for the testing sets. In our experiments, we believe the query-independent experiments conducted in cross-validated fashion with AUC as a performance metric are more general than averages over sets of example queries.

A limitation of our study is that we do not compare the machine learning models to PageRank. Computation of PageRank for MEDLINE requires access to the *complete* proprietary citation database of ISI which is not available to the public or to the research community (with the exception of [9] to the best of our knowledge). The problem is alleviated to a large extent since [9] established that citation count is better than PageRank so by transitivity our experiments suggest that GSS machine learning is superior to PageRank as well. However, the present study did not produce the data that would directly support a similar argument for the ACPJ gold standards, and when the full citation data becomes available to the research community it will be interesting to produce comparisons of PageRank to GSS models built for a variety of tasks and gold standards.

The study in the present paper is furthermore limited in the use of three tasks and corresponding gold standards out of many possible ones. Several more studies with other tasks, specialties of medicine, time horizons, and gold standard corpora/criteria will be needed before the relative value of focused filters versus citation metrics is entirely understood.

Discussion and Conclusions

An article may cite another article for a variety of reasons: authors may cite articles to acknowledge prior work, identify methodology, provide background reading, correct or criticize, substantiate claims, alert readers to forthcoming work, authenticate data, identify original publication of a term or concept, disclaim work of others, or dispute priority claims [39]. In addition, the citing paper may be a comprehensive review that attempts to cite most recent papers on the topic, the article reviewers may have recommended that the citation be included, the cited article may be a highly controversial or fashionable one, etc. An article citation thus may or may not endorse a cited article. The lack of an unambiguous connection between citation, context of use, manner of use, and/or endorsement prevents citation count from being a single effective measure of inclusion in an “importance” bibliography. More generally stated, the conceivable reasons for citation are so numerous that it is unrealistic to believe that citation conveys just one semantic interpretation. Instead citation metrics are a superimposition of a vast array of semantically distinct reasons to acknowledge an existing article. It follows that any specific set of criteria cannot be captured by a few general citation metrics and only focused filtering mechanisms, if attainable, would be able to identify articles satisfying the specific criteria in question.

Another limitation of citation metrics is that they assume that the frequency of citations is uniform across all topics. This assumption is clearly not true for all topics in biomedicine. For example, the total number of citations using the query “breast cancer” in Pubmed returns 141,704 citations whereas the query “osteosarcomas” returns 15,904 articles (executed on 11/15/2005) [40]. Thus even the highest ranking article in osteosarcomas by citation count may not rank comparably to articles at lower ranks within breast cancer.

We also note that citation metrics are not only limited by their lack of focus, but, in general, they are not available until several years have passed, which reduces their usefulness for assessing cutting-edge articles. Since predicting future citation count is an open and unsolved problem in pattern recognition so far, it follows that citation metrics are not only too non-specific but also unavailable when needed the most (i.e., for articles published in recent years).

How feasible and practical is it to built GSS focused filters for identifying high quality articles? Several examples of recent research has provided evidence that construction of focused filters is feasible and practical using both manual and machine learning approaches for non-trivial sets of criteria [1, 2, 4, 41].

We observe that the SSOAB machine learning models’ discriminatory performance as measured by the AUC indicates, in addition to theoretical interest, promising potential for practical application. As an indicative example, consider a query (in the domain -surgical oncology, internal medicine, etc- for which the model is trained), that returns 1,000 MEDLINE documents, a number which by any standard is very difficult to check manually. A reasonable prior for high-quality articles as informed by the literature on quality corpora is about 5% [10], which means that there are 50 important documents in the 1000 relevant ones[3]. By applying the SSOAB machine learning model threshold that corresponds to the ROC point with sensitivity and specificity of 85% and 85% (AUC of 0.89 from experiment 1) correspondingly, a system built around these models would select 186 documents of which 43 are true positive and 143 false positive. This filtered document set is more manageable by manual inspection. Additionally, further improvements of AUC would improve identification of high quality articles. For example, with the same scenario, at 90% sensitivity and 93% specificity (AUC of 0.97 for ACPJ treatment category in experiment 4), 112 articles would be returned with 45 true positives (out of 50) and 67 false positives. In relevance queries that return fewer documents to begin with (e.g., 200 documents) a user might select a point on the ACPJ treatment ROC curve that has 99% sensitivity and 70% specificity which would return all 10 true positives and 57 false positives.

In conclusion, whereas the appeal of "one metric fits all needs" is indeed powerful, and citation counts are fairly easy to obtain, the experiments we present together with the inherent theoretical limitations of citation metrics we discussed demonstrate that context-free citation approaches are inferior to focused filters built for specific tasks and gold standards. Furthermore, including citation metrics as predictors does not give extra advantages to the focused filters. We propose that a divide-and-conquer approach that

[3] Caveat in the above example scenario: the proportion of positive documents may vary between query results. The overall prior of positives mentioned corresponds to average performance over all queries.

uses GSS focused filters for well-defined queries, contexts of use, and quality criteria as more likely to be successful than context-free citation metrics.

Acknowledgements

The first author acknowledges support from NLM grant LM007948-02 and the Medical Library Association's Donald Lindberg Research Fellowship. The last author acknowledges support from grant LM007948-01. Conversations and prior joint work with Drs. Elmer Bernstam and Bill Hersh provided the stimulus for formulating the hypothesis of the present paper.

References

1. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text Categorization Models for High Quality Article Retrieval in Internal Medicine. J Amer Med Inform Assoc. 2005;12(2):207-216.

2. Haynes B, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing Optimal Search Strategies for Detecting Sound Clinical Studies in MEDLINE. J Amer Med Inform Assoc. 1994;1(6):447-458.

3. Wilczynski N, Haynes B. Optimal Search Strategies for Detecting Clinically Sounds Prognostic Studies in EMBASE. J Amer Med Inform Assoc. Jul/Aug 2005;12(4):481-485.

4. Duda S, Aliferis CF, Miller RA, Statnikov A, Johnson KB. Extracting Drug-Drug Interaction Articles from MEDLINE to Improve the Content of Drug Databases. In: AMIA Symposium; 2005; Washington, D.C.

5. Garfield E. The Meaning of the Impact Factor. International Journal of Clinical and Health Psychology 2003;3(2):363-369.

6. Garfield E, Welljams-Dorof A. Citation data: their use as quantitative indicators for science and technology evaluation and policy-making. Science and Public Policy 1992;19(5):321-327.

7. Page L, Brin S, Motwani R, Winograd T. PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1998.

8. Kleinberg. Authoritative Sources in a Hyperlinked Environment. Proceedings of the ACM-SIAM Symposium on Discrete Algorithms. 1997.

9. Bernstam EV, Herskovic JR, Aphinyanaphongs Y, Aliferis CF, Sriram MG, Hersh WR. Using Citation Data to Improve Retrieval from MEDLINE. J Amer Med Inform Assoc. (e-pub ahead of print). October 14, 2005:doi: 10.1197/jamia.M1749.

10. ACP_Journal. Purpose and Procedure. ACP Journal 1999;131(1):A-15 - A-16.

11. SSOAB. (Accessed: 12-05-2005), http://www.surgonc.org.

12. Web Of Science. (Accessed: 12-05-2005), http://www.isinet.com/products/citation/wos.

13. Journal Citation Reports. (Accessed: 12-05-2005), http://www.isinet.com/products/evaltools/jcr.

14. Salton G, Buckley C. Term weighting approaches in automatic retrieval. Information Processing and Management 1988;24(5):513-523.

15. Porter M. An algorithm for suffix stripping. Program 1980;14(3):130-137.

16. MEDLINE Stopwords. (Accessed: 12-05-2005), http://biolib.princeton.edu/instruct/MedSW.html.

17. Leopold E, Kindermann J. Text Categorization with Support Vector Machines. How to Represent Texts In Input Space? Machine Learning 2002;46:423-444.

18. Vapnik V. Statistical Learning Theory. New York: Wiley; 1998.

19. Joachims T. Learning to Classify Text Using Support Vector Machines: Kluwer; 2002.

20. LIBSVM: a library for support vector machines. (Accessed: 12-05-2005), http://www.csie.ntu.edu.tw/~cjlin/libsvm.

21. Dumais S, Platt J, Heckerman D, Sahami M. Inductive learning algorithms and representations for text categorization. In: Proceedings of ACM-CIKM98; 1998 November.

22. Hsu C-W, Chang C, Lin C. A practical guide to support vector classification. Technical Report 2005. Available on-line at http://www.csie.ntu.edu.tw/~cjlin/libsvm. Accessed 12-05-2005.

23. Yang Y, Liu X. A Re-Examination of Text Categorization Methods. In: 22 Annual ACM Conference on Research and Development in Information Retrieval.; 1999; Berkeley, CA: ACM Press.

24. Sun A, Lim E, Ng W. Hierarchical Text Classification and Evaluation. In: ICDM; 2001.

25. Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. New York: ACM Press; 1999.

26. Fawcett T. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Technical Report: HP Labs.; 2003. Report No.: HPL-2003-4.

27. Hand DJ, Till RJ. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning 2001;45:171-186.

28. Provost F, Fawcett T, Kohavi R. The Case Against Accuracy Estimation for Comparing Induction Algorithms. In: ICML-98 (15th International Conference on Machine Learning); 1998.

29. Pagano M, al e. Principles of Biostatistics. Australia: Duxbury Thompson Learning; 2000.

30. Weiss S, Kulikowski CA. Computer Systems that Learn. San Mateo, CA: USA Morgan Kauffman; 1991.

31. Scheffer T. Error estimation and model selection. Technischen Universit at Berlin; 1999.

32. Dudoit S, Van Der Laan MJ. Asymptotics of cross-validated risk estimation in model selection and performance assessment. Working Paper: U.C. Berkeley Division of Biostatistics; 2003 February 5. Report No.: 126.

33. Aliferis C, Tsamardinos I, Statnikov A. HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection. In: Proceedings AMIA Symposium; 2003; Washington DC.

34. Tsamardinos I, Aliferis C. Towards principled feature selection: relevancy, filters, and wrappers. In: AI and Statistics; 2003.

35. Aphinyanaphongs Y, Aliferis CF. Learning Boolean Queries for Article Quality Filtering. In: MEDINFO; 2004; San Francisco, CA.

36. Aliferis CF, Tsamardinos I, Statnikov A, Brown LE. Causal Explorer: A Causal Probabilistic Network Learning Toolkit for Biomedical Discovery. METMBS 2003:371-376.

37. Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 2002;46:389-422.

38. Delong E, Delong D, Clarke-Pearson D. Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-45.

39. Garfield E, editor. Can citation indexing be automated? Washington, DC: National Bureau of Standards; 1965.

40. PubMed. (Accessed: 12-05-2005), http://www.ncbi.nlm.nih.gov/PubMed/.

41. Jenkins M. Evaluation of Methodological Search Filters - A review. Health Information and Libraries Journal 2004;21:148-163.

Appendix : HITS Curves and Precision – Recall Curves from Cross-Validated Design.

The HITS and precision-recall curves were generated using the cross-validated design. Both curves were generated independently within each fold of the cross-validated design, and an average composite curve for both metrics was generated as an average over the curves from each fold (Figure III-1 and Figure III-2).

These curves should not be compared to the HITS and precision-recall curves generated in [9] due to differences in the experimental datasets (see Study Limitations).

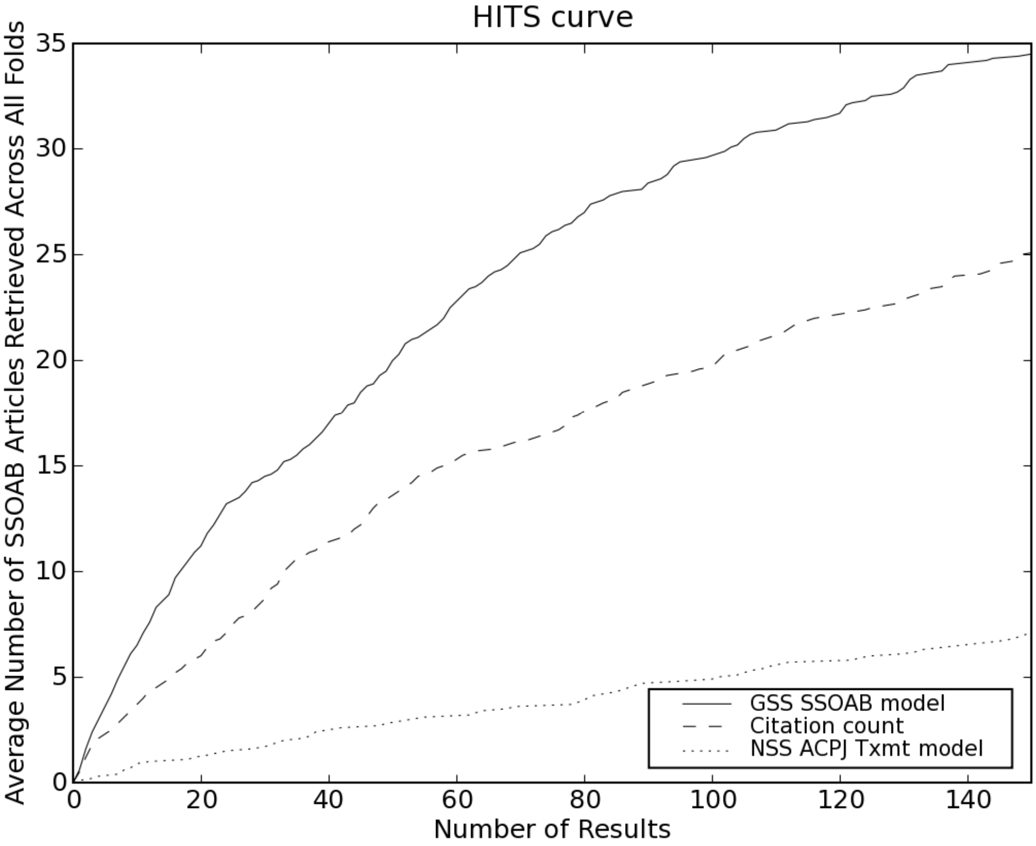


Figure III-1 – Average HITS curves used on SSOAB corpus. The GSS SSOAB model returns the most documents in the first 150 articles. Citation count and the NSS ACPJ Txmt Model applied to the SSOAB corpus return fewer documents in the top 150 returns. The SSOAB corpus was composed of 431 positives and 7,379 negatives. Each curve was an average across all folds.

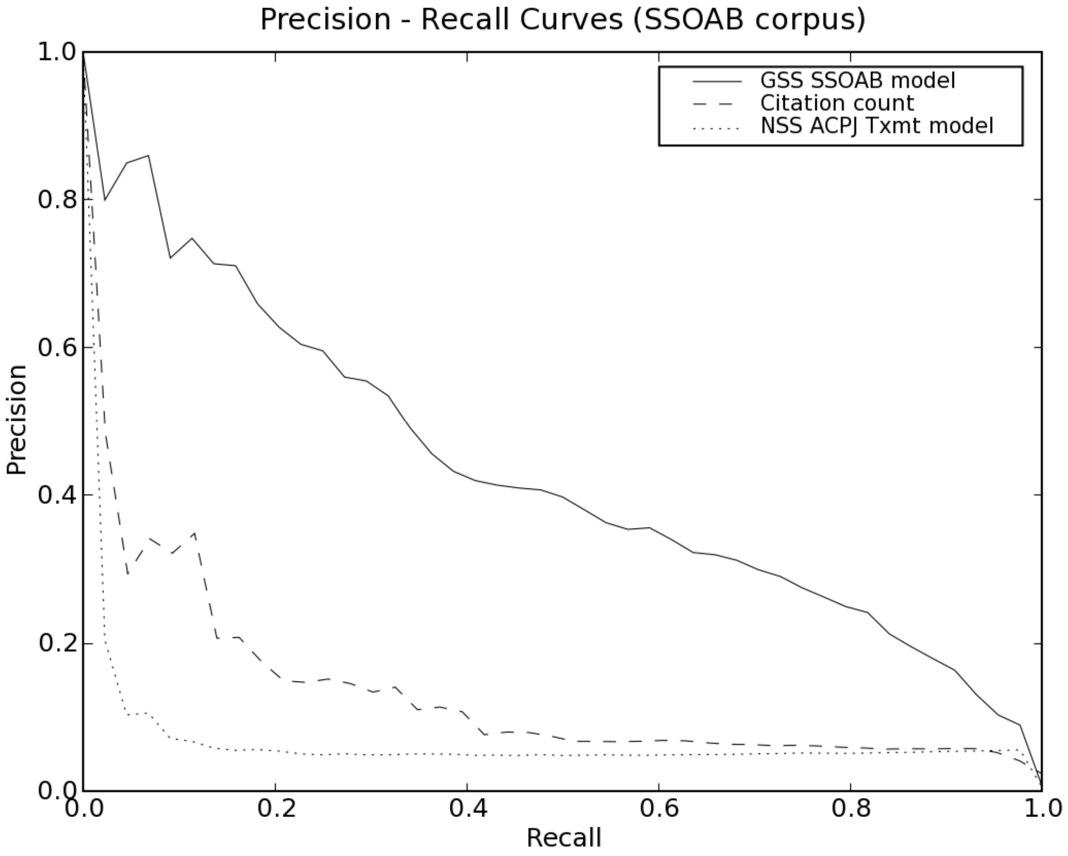


Figure III-2 – Average precision-recall curves used on SSOAB corpus. The GSS SSOAB model returns the best performing precision-recall curve. Citation count and the NSS ACPJ Txmt Model have curves below, thus performing lower than, the GSS SSOAB model. The SSOAB corpus was composed of 431 positives and 7,379 negatives. Each curve was an average across all folds.

A Comparison of Web Hyperlinks to Machine Learning Filters for the Identification of High Quality MEDLINE Documents

Abstract

Mining web hyperlinks between web pages has formed the basis for the commercial success of search engines such as Google and Yahoo. We hypothesized that mining web hyperlinks and web pages on the internet may identify high quality MEDLINE articles. Our assumption for hyperlink analysis to be successful in identifying quality MEDLINE articles is that medical sources on the web are more likely to link and cite higher quality medical literature. We built a gold standard database based on article selections by the ACP Journal Club in the treatment, etiology, diagnosis, and prognosis content categories. We ranked these articles using Google PageRank, Yahoo WebRanks, 2005 impact factors, total web page hit count for each article, bibliometric citation count, and machine learning filter models. We generated receiver operating curves and calculated area under the curves as a measure of discriminatory power for identifying high quality articles using each method. The machine learning filter models had superior performance of 0.95 average areas under the curve across the 4 categories. Bibliometric citation count, total web page hit count, impact factor, Google, and Yahoo had average area under the curves of 0.68, 0.60, 0.58, 0.53, and 0.49 respectively. Bibliometric citation measures and web-based ranking measures are *not* effective in identifying high quality articles in the medical literature. Machine learning filter models have superior performance in identifying high quality articles from an ACP Journal Club gold standard in 4 content categories.

Introduction

An article or journal receiving a high number of citations is an indicator of impact and possibly quality. Since the 1960s, the calculation of an "impact factor" of a journal has been based on counting the number of citations to a journal in a fixed time period as a measure of the impact of the journal. Impact factors have proven to be valuable, successful, and often controversial in identifying impact for journals and research in the literature [1-3].

The web analog to bibliometric citation count is hyperlinks. Search engines such as Google and Yahoo use ranking algorithms which rely heavily on the hyperlinked structure of the web. The basic premise of these linked based algorithms is that the number and quality of hyperlinks to a web site determine its importance (reported as a score). The number of hyperlinks is defined as the raw number of hyperlinks to the page. Quality is defined as the importance of the web page with the hyperlink pointing to the page. Thus, a web page gains "importance" in two ways: first, by having many links pointing to the web site, and second, by having high quality hyperlinks from web sites such as Yahoo pointing to the web page [4].

An interesting hypothetical question is whether the success of these link based measures would carry over to identifying MEDLINE articles using web based hyperlinks. Our assumption for hyperlink analysis to be successful in identifying quality MEDLINE articles is that medical sources on the web are more likely to link and cite higher quality medical literature to substantiate claims or otherwise. With the ease of creating a hyperlink, and the volume of hyperlinks available, it seems reasonable to assume that hyperlink measures could identify quality articles in the primary literature.

Answering this question has practical significance. First, if high quality MEDLINE articles are cited more often and with higher rank with Google PageRank or Yahoo Webrank on the web than low quality articles, general search engines may be a viable means for ranking medical literature. Second, if identifying high quality articles is marginal, we provide evidence that health professionals should take caution in using general search engines to identify evidence from the medical literature.

In this study, we explored the usefulness of web hyperlinks in identifying high quality articles in MEDLINE. We compared rankings by total number of pages citing a reference, rankings by Yahoo, rankings by Google, rankings by bibliometric citation count, and rankings by machine learning filters for the same task. We hypothesize that web citation metrics have marginal performance in identifying high quality MEDLINE articles and do not perform as well as machine learning filter models for the same task because of the unfocused nature of hyperlinks and web pages on the internet.

Background

In [5], Bernstam and colleagues compared bibliometric citation count to machine learning filters built for a different task to identify articles selected for a surgical oncology society bibliography. Bernstam and colleagues concluded that bibliometric citation count had superior discriminatory performance over impact factors, the Pagerank algorithm (as applied to the bibliometric citation graph), Pubmed, sensitive clinical query filters of Pubmed, specific clinical query filters of Pubmed, and EBMSearch (a search engine implementing machine learning filters built for a different task). In subsequent work, we explored the use of machine learning filters built specifically for the surgical oncology task [6]. We showed that machine learning filters built specifically for the task had superior discriminatory performance over bibliometric citation count, the EBMSearch filters built for a different, treatment specific task, impact factors in 2001, and impact factors in 2005.

As far as we know, there is no research that uses web citation metrics to identify high quality MEDLINE articles. Instead, researchers focused on high quality web pages. They used web citation metrics, specifically scaled PageRank scores between 0 and 10 as reported by the Google browser toolbar [7], to identify high quality web pages[4].

Fricke and Fallis [8] evaluated PageRank score as one indicator of quality for 116 web sites about carpal tunnel syndrome. They made a strong statement that web sites with a Google PageRank score greater than 5 had good quality of information about carpal tunnel syndrome. Of the 57 inaccurate web sites, 29 had PageRank scores greater than 5 while of the 59 accurate web sites, 41 had PageRank scores greater than 5. They

[4] Versions of the Google toolbar as if this publication do not display Pagerank scores for web pages.

concluded that, as a univariate measure, PageRank scores select more accurate websites at this threshold.

However, the conclusion is problematic. First, the threshold is too low. Users typically do not look beyond the top 10 or 20 results of a search engine [9]. A practical question is that in the range of web sites selected by users, is PageRank score an indicator of quality. Second, the use of PageRank score is not inherently useful for discrimination or helping users to avoid inaccurate or poor information. The results imply that even with the 70 web sites with high PageRank, 29 of them will have inaccurate information. Unfortunately, Fallis and Fricke did not calculate a correlation between PageRank score and quality.

Griffiths [10] evaluated PageRank scores with evidence based quality scores for depression websites. The authors obtained Google PageRank scores for 24 depression websites from the DMOZ Open Directory Project website. Two health professional raters assigned an evidence based quality score to each site. PageRank scores correlated weakly (r = 0.61, P=0.002) with the evidence based quality scores. Despite this, the authors concluded that as a screening tool, PageRank may be an appropriate technique to exclude low quality sites.

Tang, Craswell, and Hawking [11] compared Google results with a domain-specific search engine for depression. They found that of a 101 selected queries, Google returned more relevant results, but at the expense of quality. Of the 50 treatment related queries, Google returned 70 pages of which 19 strongly disagreed with the scientific evidence. These authors concluded that a tension exists between relevance and quality, and indexing more pages may give a greater number of results, but selective inclusion can give better quality.

In summary, Google PageRank seems a valid, but rather weak, baseline to compare against any pattern recognition based, automated method to identify high quality websites. Yahoo Webrank scores have not been studied in the context of high quality web pages.

Methods

A. Google PageRank and Yahoo WebRank

Google PageRank and Yahoo WebRank algorithms are proprietary. Google relies on over a hundred different factors in ranking web pages [12]. In general, both rely on combinations of anchor text, meta-tags, traffic patterns, and hyperlinks for ranking web pages. For a more thorough description of the core Google PageRank algorithm, we refer the reader to [4]. We could not find any resources on the Yahoo Webranks algorithm describing the ranking algorithm.

The raw scores from Google or Yahoo are not available. Google provides a proxy to the score output from the PageRank algorithm through the Google browser toolbar [7]. Though the exact scaling is not known, search engine optimizations analysts have suggested that the toolbar logarithmically scales the output of the PageRank algorithm to integer values between 0 and 10. In early 2005, Yahoo also provided a 0-10 integer WebRank score through the Yahoo browser toolbar [13]. Since then, the score has been removed from the toolbar, and the WebRank scores are not available either through the toolbar or the application programming interface [14].

Since the raw scores are not available, we used paired comparisons and a mergesort algorithm [15] to rank the articles for both Yahoo and Google. To sort a list of articles, one article must rank higher than another article. An article ranks higher than another article if it appears in the results list of Google or Yahoo first. To make the comparison, we queried the respective search engines with the titles of two articles in quotes "OR"ed together. Then beginning with the first returned web page, we parsed each web page and matched the title and the first author's last name in the web page. Whichever article title and first author's last name matched first, gets a higher ranking than the corresponding paired citation. Through series of these comparisons, we eventually sorted the list of articles. The resulting ranks should be equivalent to the exact rankings if it was possible to query the search engine with a large Boolean "OR" composed of all titles and first author's last name of all articles. The paired comparisons and sorts were run over a time period of several months spanning September 2006 to January 2007.

B. Web Page Hit Count

We constructed a Google query using the first author's last name and words of the title in quotes. For each query, we obtained the total number of web pages returned citing the reference. The total results count is available through the Google search api. I obtained the web page hit counts in January of 2007.

C. Classification Methods

In our experiments, we employed Support Vector Machine (SVM) classification algorithms. These methods calculate a maximal margin hyperplane separating two or more classes of the data. To accomplish the separation, the data is mapped to a higher dimensional space by means of a kernel function, where a separating hyperplane is found by solving a constrained quadratic optimization problem [16]. For several published text categorization tasks, SVMs have had superior classification performance compared to other methods [17, 18], and this motivated our use of them. We used an SVM classifier implemented in libSVM v2.8 [19] with a polynomial kernel. We optimized the SVM penalty parameter C over the range {0.1, 1, 10} with imbalanced costs applied to each class proportional to the priors in the data [20], and degree d of the polynomial kernel over the range. Since theoretical literature on domain characteristics as it relates to optimal parameter selection is not yet developed , the ranges of costs and degrees for optimization were chosen based on previous empirical studies [17, 18, 21]. Different combinations of costs and degrees were exhaustively evaluated by cross-validation, and the best performing model was selected for the final application of the SVM classifier (see section on Performance Estimation and Model Selection)

D. Bibliometric Citation Count

Bibliometric citation count is the number of publications citing an article. We downloaded citation counts from the Web of Science [22] using a screen scraping interface coded in Python [23]. The screen scraper established an http connection to the Web of Science servers and navigated through several GET and POST requests to identify an article and parse out the number of cited articles. We manually determined the number of cited articles for any articles in the few cases where the automated method failed. We obtained citation counts of articles in the ACPJ gold standards [17, 24] on August 2002. These collection dates allowed approximately 2.5-3 years for citations to accumulate in each respective gold standard.

A relatively small number of articles did not have a citation count since the corresponding journals were not followed by the Web of Science. For the ACPJ gold standard, we obtained 13,279 citations that had counts out of 15,786 citations in the gold standard. Articles without citation counts were excluded from the study.

E. Impact Factor

An impact factor of a journal is the average number of citations an article published in the journal receives in two years [25]. For example, the 2004 impact factor for journal X would be the number of citations received by articles published in X within 2002-03, divided by the total number of published articles in X within 2002-03. We obtained impact factors from the Web of Science for 2005.

F. Performance Metrics

We used receiver operating characteristic (ROC) curves to analyze performance of the classification algorithms [26]. Intuitively, these curves depict the tradeoff between correct and incorrect classification. The ROC curve is plotted in the dimensions of (1-specificity) and sensitivity. In inspecting ROC curves, the point (1,0) (i.e. specificity=1 and sensitivity=0) corresponds to a classifier where all examples are classified as negative, the point (1,1) where all examples are classified as positive, and the coordinate (0,1) is where all examples are classified perfectly. The closer a point on the ROC curve is to the (0,1) coordinate, the better performance the classifier has, assuming that false positives and false negatives have the same cost.

A composite measure of ROC performance is typically reported as the area under the ROC curve (AUC) [27]. Areas range from 0 to 1 with 1 being perfect classification, 0.5 being random classification performance, and 0 being an inverse classification with all true positives classified as negatives and all true negatives classified as positives [26]. The AUC performance metric has the important property of being invariant to class distribution. Class invariance is essential for this domain since we often have many more negative documents than positive ones. Relying on performance measures that are sensitive to class distributions may be a misleading measure of discriminatory performance. For example, we would not use accuracy (defined as the proportion of correct classifications over all classifications) as a performance measure, because excellent accuracy can be achieved in extremely skewed distributions by classifying all documents as belonging to the most prevalent class [28].

G. Document Representation and Pre-processing for Machine Learning

The conversion of documents to a format suitable for the machine learning algorithms followed the procedures in [17]. The articles in the ACPJ selected journals were cross-referenced in PubMed, and the title, abstract, journal, and MeSH terms were extracted. We represented each document as a set of terms for the learning algorithms [29]. We stemmed each term [30], removed “stopword” terms [31], and any terms occurring in less than 5 documents. Very infrequent terms are difficult to assess statistically and may affect negatively the generalization of the classification models. Selected terms were further encoded as weighted features using a log frequency with redundancy scheme for all documents [32]. The ACPJ etiology and treatment collection contained, after imputation of citation counts, 15,786 articles with abstracts and citation counts represented by 28,229 features including citation count.

H. Performance Estimation and Model Selection

We used 5-fold cross-validation to estimate the performance of the learning algorithms [33]. This procedure first divided the data randomly into 5 non-overlapping subsets of documents (subject to the constraint that the proportion of positive and negative documents in the full dataset is preserved for each subset). Next, the following was repeated 5 times: we used one subset of documents for testing (the “original testing set”) and the remaining four subsets for training (the “original training set”) of the classifier. The average performance over 5 original testing sets is reported.

In order to optimize parameters of the learner (e.g., SVM algorithms), we used another “nested” loop of cross-validation by further splitting each of the 5 original training sets into smaller training sets and validation sets. For each combination of learner parameters, we obtained cross-validation performance and selected the best performing parameters inside this inner loop of cross-validation. We next built a model with the best parameters on the original training set and applied this model to the original testing set. Details about the “nested cross-validation” procedure can be found in [34, 35]. Notice that the final performance estimate obtained by this procedure will be unbiased because each original testing set is used only once to estimate performance of a single model that was built by using training data exclusively.

I. Gold Standard Construction

1. ACPJ Journal Club

The ACP Journal Club is a highly-rated meta-publication. Every month expert clinicians review a broad set of journals in internal medicine, and select articles in these journals according to specific criteria in the content areas of treatment, diagnosis, etiology, prognosis, quality improvement, clinical prediction guide, and economics. Selected articles are further subdivided into articles that are summarized and abstracted by the ACP because of their "clinical importance", and those that are only cited because they meet all the quality selection criteria but may not pertain to vitally "important clinical areas". For the purposes of the present study, articles that are abstracted or cited by the ACP are considered positive instances and all other articles in the *same* journals to be negative. The criteria for inclusion in ACPJ can be found in [24].

We used for the present study a modified version of the ACP Journal corpus. The Google [5] and Yahoo programming interfaces [36, 37] only allow a 1000 daily requests to the search interface. Due to this request limitation and the paired sorting algorithm which guarantees sort in O(n log n), we limit the total number of articles in the data sets to 1000. The distribution of positives and negatives is illustrated in Table III-7.

[5] Direct programmatic access to Google's search results is no longer available. The search results of the current api must be screen scrapped to obtain the same search results.

Table III-7 - Positives and negatives distribution in each category. The sizes were limited due to Google and Yahoo API restrictions.

Category	**Positives**	**Negatives**
Treatment	297	750
Etiology	169	855
Prognosis	21	902
Diagnosis	29	901

J. Experimental Procedure

In each category, we created a dataset composed of the positives and negatives as shown in Table III-7. We ranked the articles in each category by Google PageRank, Yahoo WebRank, impact factor, web page hit count, bibliometric citation count, and machine learning filter model score.

We ranked the articles by Google PageRank and Yahoo WebRank by paired article comparisons as described in the Methods Section. Articles were sorted by rank, and an area under the curve was calculated. The paired comparisons and sorts were run over a time period of several months spanning September 2006 to January 2007.

We ranked articles from 2005 impact factors obtained from the Web of Science. Articles were ranked by the impact factor of the originating journal. Area under the curve was calculated.

We ranked articles by web page hit count by querying Google with the last name of the first author and the title of the article in quotes. We used the estimated total result count in the Google result set as a measure of the number of web pages citing the selected article. The articles were ranked according to the total hit count and an area under the curve was calculated. The web page hit counts were obtained in January of 2007.

We ranked articles by bibliometric citation count by screen scraping the Web of Science in August 2002 as described in the Methods Section. Articles were ranked according to total bibliometric citation count and an area under the curve was calculated.

For the machine learning models, we used 5 fold cross validation to estimate the area under the curve performance for each category as explained in the Methods section.

Results

The area under the curves for each ranking method are shown in Table III-8 and the average area under the curve in Table III-9.

Table III-8 - Area under the ROC curve for each ranking method. *- area under the curve below 0.5 indicates that the method ranks articles in reverse order. Thus reversing the ranking would rank the positives higher than the negatives.

Method	**Treatment**	**Etiology**	**Prognosis**	**Diagnosis**
Google PageRank	0.54	0.54	0.43*	0.46*
Yahoo WebRanks	0.56	0.49*	0.52	0.52
Impact Factor 2005	0.67	0.62	0.51	0.52
Web page hit count	0.63	0.63	0.58	0.57
Bibliometric Citation Count	0.76	0.69	0.67	0.60
Machine Learning Filter Models	**0.96**	**0.95**	**0.95**	**0.95**

Table III-9 - Average area under the ROC curves for each method.

Method	**Average area under the curve across 4 categories.**
Google PageRank	0.53
Yahoo WebRanks	0.49
Impact Factor 2005	0.58
Web page hit count	0.60
Bibliometric Citation Count	0.68
Machine Learning Filter Models	**0.95**

Ranking by Google Pagerank and Yahoo Webranks have the lowest discriminatory power compared to impact factor, web page hit count, bibliometric citation count, and the machine learning filter models. In the diagnostic and prognostic categories, the article rankings by Google are reversed in placing negatives above positives. Impact factor has limited ability to discriminate quality of articles. Web page hit count has some discriminatory power with bibliometric citation count having the next highest discriminatory power. The machine learning filter models have the best discriminatory power of the compared methods upwards of 0.95 in all categories.

Study Limitations

Sample was limited for each category. Each category was limited to a 1000 articles due to the limits set by the Google and Yahoo APIs. Ranking the 15,000 or 30,000 articles through paired comparisons is possible and a point of future research.

Another limitation was that web pages citing an article were identified using queries to Google and Yahoo of the last name of the first author and the title words in quotes. We did not run formal experiments to establish how well web pages citing an article were identified using this query. In initial experiments, we observed that queries with title words in quotes would rarely return web pages that did not cite the article. In subsequent experiments, manual ad-hoc inspection of the top 10 results of 50 random queries confirmed this observation that the queries identified web pages citing an article with high sensitivity and specificity.

We did not compare to Google Scholar [38]. Google does not provide an API to this site. Though the exact inner workings of Google Scholar are not available, the web site states “Google Scholar aims to sort articles the way researchers do, weighing the full text of each article, the author, the publication in which the article appears, and how often the piece has been cited in other scholarly research. The most relevant results will always appear on the front page.” How this translates to ranking our articles is unknown and a point of future research.

Another potential limitation is that the articles selected in our dataset are from 1998-1999. It is conceivable but not likely that web based measures would not rank correctly older articles. Older articles on the web may not be cited as often or at all if the articles themselves are obsolete. A point of future research would be to replicate these experiments with a more recent corpus.

Discussion and Conclusions

Machine learning methods outperformed web and bibliometric citation measures in discriminating high quality articles. We reached a similar conclusion in prior work when comparing bibliometric citation counts to machine learning methods with the space of MEDLINE documents [6]. The same reasoning for why bibliometric citation count does not discriminate as well as machine learning algorithms also applies to web hyperlinks. Links do not necessarily confer authority to the linked page.

In prior work in using web hyperlinks to identify quality web pages, Chakrabarti observed that many links have nothing to do with the conferral of authority [39]. Some links exist purely for navigational purpose or as paid advertisements. Chakrabarti et al, hope that, in an aggregate sense, over a large enough number of links, the view of links as “conferring authority” will hold.

Aggregate links do not confer authority for identifying high quality articles in medicine. Rankings by Google and Yahoo rely on link analysis, and thus, by proxy, should rank quality articles higher. According to this gold standard, higher quality articles are not preferentially cited over lower quality articles or cited on higher quality web sites.

A potential future direction to leverage link information is to attempt to establish intent of the links. Would it be possible to determine when a link confers authority rather than other links that are navigational or advertising in nature. It may be possible that reduction of link “noise” would improve the ranks of higher quality articles.

But the assumption that higher quality articles are cited *more often* than lower quality articles does seem to hold as illustrated by the discriminatory power of the web page hit count. We made an initial assumption that secondary sources of medical information would cite higher quality articles more often than lower quality articles. The discriminatory power of web page hit count is interesting and should be explored in future work.

Another potential future direction is to consider ranking articles by the intent of the web pages they appear on. For example, we compared two articles using the paired comparison test. The top web page in the results lists is a NEJM [40] citation located on a table of contents web page on the NEJM web site. This type of citation is not an explicit endorsement and is navigational in nature. Contrast this with the second article that appears at a lower rank from emedicine.com [41]. The article appears as a citation from a secondary source written by the emedicine authors. In this example, the second article should rank higher because it is on a web page that endorses the article when compared to the first article where the web site ranks highly because it is from a more popular source.

In preliminary experiments, we made an attempt to rank articles by determining the intent of the web page. We excluded pages that were navigational or contained the abstract of the article (as articles that contained the abstract were likely to be publisher pages listing the abstract or sites such as Pubmed that list the abstract of the article). In one class, using these general heuristics, we were able to improve the area under the curve from 0.54 to 0.71 in treatment. Further improvements are likely if it is possible to build filters that may determine page intent. Whether these results generalize to other content areas are unknown and an area for future research.

Ultimately coming up with a way to measure article quality through links or web page endorsement is important. A limitation of the machine learning filter model is that a explicit gold standard must be created to build the model. In contrast, it may be possible to use web links to identify articles without an explicit gold standard. Though the question does remain as to what is being ranked highly if we use web links. If we could count links that "conferred authority" only, we would find literature that was cited by secondary sources, and the articles that rank highly are articles that are cited more often by these secondary sources. A method such as this may prove to be valuable for ranking the literature. In other words, a hypothesis to be tested is that ranking the literature by how often the myriad of secondary sources cite the individual articles may be effective.

The question of why articles rank highly on the web is important. Articles that rank highly seem to follow popularity trends. Articles published in the NEJM [40] or JAMA [42] will rank higher than articles published in other journals because the websites and journals are highly cited and linked to on the web in general. The distinction between ranking by popularity and quality is not clear.

The results of this study also suggested that health professionals should use their discretion in evaluating the results of a web search. As shown in these experiments, ranking articles by Google, Yahoo, or web hyperlinks does not necessarily return the highest quality articles and instead returns articles ranked according to Pagerank or Webranks.

We also considered counting backlinks to webpages that have the individual article selected. For example, we considered counting the number of backlinks from the Pubmed page where the citation appears. The idea is that a webpage that cites the article will link to a version of the article containing the abstract either on the Pubmed site or a publisher's website. Unfortunately, at the time of this writing, the Google and Yahoo APIs do not provide complete lists of backlinks to individual web pages. We suspect that counting web page hit counts is a valid proxy for counting backlinks and is a more accurate measure of citation by web pages since most pages may not link to Pubmed or publisher versions of the citation.

Finally, the machine learning filter models perform best when compared to these citation metrics. The machine learning filters are focused on their intent and will identify articles that match the criteria of the gold standard. The superior performance of these filters to other citation measures is supported by our previous evaluative studies [6, 17].

Conclusions

We have compared machine learning filter models to Google PageRank, Yahoo Webranks, impact factor, bibliometric citation count, and web page hit count for discriminating high quality articles from an ACP Journal Club gold standard. The machine learning filter models had superior performance in identifying high quality articles from an ACP Journal Club gold standard in the categories of prognosis, diagnosis, treatment, and etiology.

References

[1] Seglen P. Why the impact factor of journals should not be used for evaluating research. BMJ. 1997 Feb 15;314(7079):498-502.

[2] Monastersky R. The Number That's Devouring Science. The Chronicle of Higher Education. 2005 October 15.

[3] Not-so-deep-impact. Nature. 2005 June 23;435:1003-4.

[4] Page L, Brin S, Motwani R, Winograd T. PageRank Citation Ranking: Bringing Order to the Web. Technical Report. 1998.

[5] Bernstam EV, Herskovic JR, Aphinyanaphongs Y, Aliferis CF, Sriram MG, Hersh WR. Using citation data to improve retrieval from MEDLINE. J Am Med Inform Assoc. 2006 Jan-Feb;13(1):96-105.

[6] Aphinyanaphongs Y, Statnikov A, Aliferis CF. A Comparison of Citation Metrics to Machine Learning Filters for the Identification of High Quality MEDLINE Documents. J Am Med Inform Assoc. 2006 Apr 18.

[7] Google Toolbar. [cited 2007 6-20]; Available from: toolbar.google.com

[8] Fricke M, Fallis D, Jones M, Luszko GM. Consumer health information on the Internet about carpal tunnel syndrome: indicators of accuracy. Am J Med. 2005 Feb;118(2):168-74.

[9] Jansen B, Spink A, Saracevic T. Real life, real users, and real needs: A study and analysis of user queries on the web. Info Processing and Management. 2000;36(2):207-27.

[10] Griffiths KM, Tang TT, Hawking D, Christensen H. Automated assessment of the quality of depression websites. J Med Internet Res. 2005;7(5):e59.

[11] Tang TT, Craswell N, Hawking D, Griffiths KM, Christensen H. Quality and Relevance of Domain-specific Search: A Case Study in Mental Health. Information Retrieval. 2006;9(2):207-25.

[12] Blachman N. How Google Works. 2007 Feb 2 [cited 2007 6-20]; Available from: http://www.googleguide.com/google_works.html

[13] Yahoo Toolbar. [cited 2007 6-18]; Available from: http://toolbar.yahoo.com

[14] Yahoo Developer Network. [cited 2007 6-20]; Available from: http://developer.yahoo.com

[15] Sedgewick B. Algorithms in C++: Parts 1 - 4. Reading, MA: Addison-Wesley 1998.

[16] Vapnik V. Statistical Learning Theory. New York: Wiley 1998.

[17] Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text Categorization Models for High Quality Article Retrieval in Internal Medicine. J Amer Med Inform Assoc. 2005;12(2):207-16.

[18] Joachims T. Learning to Classify Text Using Support Vector Machines: Kluwer 2002.

[19] Chang C, C. L. LIBSVM: a library for support vector machines. 12-05-2005 [cited; Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[20] Morik K, Brockhausen P, Joachims T. Combining Statistical Learning with a Knowledge Based Approach: A Case Study in Intensive Care Monitoring. 16th International Conference on Machine Learning (ICML 1999); 1999; 1999.

[21] Dumais S, Platt J, Heckerman D, Sahami M. Inductive learning algorithms and representations for text categorization. Proceedings of ACM-CIKM98; 1998 November; 1998.

[22] Web Of Science. 12-05-2005 [cited 12-5-2005.]; Available from: http://www.isinet.com/products/citation/wos

[23] Python. [cited 2007 6-20]; Available from: http://www.python.org

[24] ACP_Journal. Purpose and Procedure. ACP Journal. 1999 July/August;131(1):A-15 - A-6.

[25] Journal Citation Reports. 12-05-2005 [cited; Available from: http://www.isinet.com/products/evaltools/jcr

[26] Fawcett T. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Technical Report: HP Labs.; 2003. Report No.: HPL-2003-4.

[27] Hand DJ, Till RJ. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning. 2001;45:171-86.

[28] Provost F, Fawcett T, Kohavi R. The Case Against Accuracy Estimation for Comparing Induction Algorithms. ICML-98 (15th International Conference on Machine Learning); 1998; 1998.

[29] Salton G, Buckley C. Term weighting approaches in automatic retrieval. Information Processing and Management. 1988;24(5):513-23.

[30] Porter M. An algorithm for suffix stripping. Program. 1980;14(3):130-7.

[31] MEDLINE Stopwords. 12-05-2005 [cited; Available from: http://biolib.princeton.edu/instruct/MedSW.html

[32] Leopold E, Kindermann J. Text Categorization with Support Vector Machines. How to Represent Texts In Input Space? Machine Learning. 2002;46:423-44.

[33] Weiss S, Kulikowski CA. Computer Systems that Learn. San Mateo, CA: USA Morgan Kauffman 1991.

[34] Scheffer T. Error estimation and model selection. Technischen Universit at Berlin; 1999.

[35] Dudoit S, Van Der Laan MJ. Asymptotics of cross-validated risk estimation in model selection and performance assessment. Working Paper: U.C. Berkeley Division of Biostatistics; 2003 February 5. Report No.: 126.

[36] Google AJAX Search API. [cited 2007 6-18]; Available from: http://code.google.com/apis/ajaxsearch/web.html

[37] Yahoo Search Web Services. [cited 2007 6-20]; Available from: http://developer.yahoo.com/search/

[38] Google Scholar. [cited 2007 6-20]; Available from: http://scholar.google.com

[39] Chakrabarti S, Dom B, Gibson D, Kleinberg J, Kumar S, Raghavan P, et al. Mining the link structure of the World Wide Web. IEEE Computer. 1999 August.

[40] New England Journal of Medicine. [cited 2007 6-20]; Available from: http://content.nejm.org/

[41] EMedicine. [cited 2007 6-20]; Available from: http://www.emedicine.com/

[42] Journal of the American Medical Association. [cited 2007 6-20]; Available from: http://jama.ama-assn.org/

CHAPTER IV

IV. EVALUATION OF GENERALIZATION OF MACHINE LEARNING MODELS

In this section, I explore the generalization of the machine learning models to identifying high quality articles. In the first set of experiments, I apply models built using a gold standard collected from 1998-2000 to a second gold standard collected in 2005. My goal was to validate the performance of the original filters on current corpora, to verify the model fitting and model error estimation procedures, and to validate consistency of the ACP Journal Club gold standard. In the second set of experiments, I explore the use of machine learning filter models in areas outside of internal medicine, in other semantic categories including clinical prediction guide, costs, and economics, and other purpose and format categories.

Prospective Validation of Text Categorization Filters for Identifying High-Quality, Content-Specific Articles in MEDLINE.

Aphinyanaphongs Y, Aliferis C. “Prospective Validation of Text Categorization Filters for Identifying High Quality, Content-Specific Articles in MEDLINE.” In: Proceedings AMIA Symposium; 2006; Washington DC.

Abstract

Finding high quality articles is increasingly difficult with the exponential growth of the medical literature. This growth requires new methods to identify high quality articles. In prior work, we introduced a machine learning method to identify high quality MEDLINE documents in internal medicine. The performance of the original filter models built with this corpus on years outside 1998-2000 was not assessed directly. Validating the performance of the original filter models on current corpora is crucial to validate them for use in current years, to verify that the model fitting and model error estimation procedures do not over-fit the models, and to validate consistency of the chosen ACPJ gold standard (i.e., that ACPJ editorial policies and criteria are stable over time). Our prospective validation results indicated that in the categories of treatment, diagnosis, prognosis, and etiology, the original machine learning filter models built from the 1998-2000 corpora maintained their discriminatory performance of 0.95, 0.97, 0.94, and 0.94 area under the curve in each respective category when applied to a 2005 corpus. The ACPJ is a stable, reliable gold standard and the machine learning methodology provides robust models and model performance estimates. Machine learning filter models built with 1998-2000 corpora can be applied to identify high quality articles in recent years.

Introduction

The purpose of a query filter is to identify medical articles that meet certain criteria (e.g., related to quality, impact, or content). Recent approaches have utilized machine learning or semi-manually constructed Boolean query based filters to pre-select articles that meet quality and content criteria [1-5]. These filters had good discriminatory performance when evaluated using cross-validation techniques [6].

Both machine learning and Boolean filters can perform much worse than expected when applied to other corpora because of two main reasons: First, it is possible for filters to be over-fitted, and second, the examples that were used to train the original filters may have a different distribution than the documents on which the filters are eventually applied [7].

Computational Learning Theory suggests that over-fitting typically occurs when filter developers fit model parameters using the training data and then estimate the future performance of the model on the same data, or when very complex models are pursued, relative to the classification function's intrinsic complexity especially in small sample learning settings (i.e., the complexity of the models considered is not tempered by the available sample and the difficulty of the learning task) [8]. Sound data modelling principles in order to avoid over-fitting include: (a) choosing model complexity and parameters that minimize both error in the training data and complexity of the model class employed; (b) estimating future (generalization) error in portions of the data reserved especially for that purpose (i.e., they are not used to fit the model) [9].

With regards to filter failure because of non-representative samples, this may occur because of small samples or very rare positive examples even if the total sample is large. In addition, non i.i.d. (independently sampled and identically distributed) sampling from the general population of documents. may lead to divergence of the training document set distribution from the application document set distribution. A particularly worrisome reason for violation of i.i.d. sampling in our context is if the gold standard for document labelling is not stable over time. For example, if the editorial policies of the ACP Journal Club changes over time, a filter built with an older editorial policy may exhibit worse performance for documents characterized as high-quality according to a more recent and thus revised editorial policy.

In this study, we address these points of failure for both Machine Learning (i.e., our own) and Boolean/semi-manual (i.e., Pubmed/Haynes et al's Clinical Query (CQ)) filters. We explore the extent of over-fitting or changes in the characteristics of the data by evaluating classification performance on articles collected independently of the original corpus. We built a machine learning filter model using a training corpus collected in one year, evaluated its performance on a prospective testing corpus collected in another year, and in the same prospective corpus, compared the machine learning filter models to the CQ filters [10] of Pubmed[6]. Thus, we have *two main hypotheses*. First, machine learning filter models built from an original corpus collected from 1998 to 2000 are able to identify high quality articles in an independently collected 2005 corpus and perform as well as estimated performance measures using cross-validation on the original corpus. Second, machine learning filter models retain their performance edge over the corresponding CQ filters in the 2005 corpus.

[6] The CQF filters are literally Boolean combinations of terms applied to a corpus. The machine learning filter models, in contrast, are not Boolean based. The machine learning filters are statistical models using all the terms in the training corpus.

Methods

Definitions

At the core of our efforts lie the selection of a rigorous quality, content gold standard and the creation of a document collection that captures this gold standard. The ACP journal club is a highly-rated meta-publication [11]. Every month experts review the best journals in internal medicine and select the best articles according to specific selection criteria in the article class areas of: *treatment, diagnosis, etiology, prognosis, quality improvement, clinical prediction guide,* and *economics*. Selected articles are further subdivided into articles that are cited and abstracted by the ACP because of their clinical importance, and those that are only cited because they meet all the selection criteria but may not pertain to vitally important clinical areas. Every article is subjected to rigorous review for inclusion [11]. By using articles abstracted and cited by the ACP as our gold standard, we capitalize on an existing high quality review.

Corpus Construction

We constructed corpora in the treatment, etiology, diagnosis, and prognosis categories spanning the time periods of July 1998 – August 1999, July 1998 – August 2000, and March 11, 2005 – August 31, 2005. From [1], we reused the two corpora built from the first two periods. For each corpus, we started with 49 journals, selected the respective time period, and collected all articles with abstracts published by these journals. We then reviewed the ACP Journal Club for at least 18 months after the specified time period for each corpus, and labeled as positive any article that was cited/ abstracted by the Journal

Club in the time period. The first corpus spanning July 1998 – August 1999 resulted in a positive/negative article distribution of 379/ 15,407 articles in treatment, and 205/ 15,581 articles in etiology. The second corpus spanning July 1998 – August 2000 resulted in a positive/negative article distribution of 74/34,864 articles in prognosis, and 102/34,836 articles in diagnosis. Refer to [1] for additional details and motivations for these constructed corpora.

We constructed the third corpus for the prospective analysis from March 11, 2005-August 31, 2005. We built the third corpus using the electronic citations available from the ACP Journal club at http://www.acpjc.org. Both articles cited and abstracted and articles cited only were available in both the print and electronic versions of the journal club. As of July 2005, the electronic version included an expanded list of articles cited only available at http://www.acpjc.org/Content/oan which we included in the independent dataset.

We covered available electronic citations in the Journal Club from July/Aug 2005 to Jan/Feb 2006 in 41 journals selected for their overlap with the 1998-2000 49 journals[7]. Because the time frame covered by the Journal Club varied from month to month, we selected 3/11/2005 as the start time period for this third corpus by averaging the earliest citation given in each journal, and the end time period of 8/31/2005 by averaging the latest citation given in each of the 41 selected journals. If no article occurs in a given journal, a date is not included in the average. Thus we selected all articles with abstracts published in 41 journals from 3/11/2005 to 8/31/2005 and identified articles cited in this time period by the ACP Journal club as positive and all others were identified as negative. This procedure resulted in a positive/negative article distribution of 351/6,921,

[7] Journal lists for both corpora are available from the authors.

47/7,601, 30/7,618, 23/7,625 in treatment, etiology, prognosis, and diagnosis respectively.

All original articles as Pubmed citations (i.e. abstracts, not full text) were downloaded with the esearch and efetch utilities available from Pubmed [12]. Each search was limited to the title of one of the journals, set to only retrieve articles during the publication period, and with the "only items with abstracts" checkbox marked. A custom parser extracted PubmedID, title, journal, abstract, publication type, and MeSH terms from the XML efetch downloads.

Article Preparation

The conversion of documents to a format suitable for the machine learning algorithm followed the procedures in [1]. The articles in the ACPJ selected journals were cross-referenced in PubMed, and the title, abstract, journal, publication type, and MeSH terms were extracted. We created two representations for each document: one for the machine learning algorithm, and one for the CQ filters.

For the machine learning algorithm, we represented each document as a set of terms for the learning algorithms [13]. We additionally stemmed each term [14], removed "stopword" terms [15], and removed any terms occurring in fewer than 5 documents. Very infrequent terms are difficult to assess statistically and may affect negatively the generalization of the classification models. Terms were further encoded as weighted features using a log frequency with redundancy scheme [16].

For the CQ filters, we represented each document as a set of terms. Words were not stemmed, but "stopwords" and infrequent terms (occurring in < 5 documents) were removed.

Statistical and Machine Learning Methods

Support Vector Machines (SVMs)

In our experiments, we employed Support Vector Machine (SVM) classification algorithms. The SVM's calculate maximal margin hyperplane(s) separating two or more classes of the data. To accomplish this, the data are mapped to a higher dimensional space by means of a kernel function, where a separating hyperplane is found by solving a constrained quadratic optimization problem [17]. SVMs have had superior text classification performance compared to other methods [1, 18], and this motivated our use of them. We used an SVM classifier implemented in libSVM v2.8 [19] with a polynomial kernel. We optimized the SVM penalty parameter C over the range {0.1, 1, 2} with imbalanced costs applied to each class proportional to the priors in the data [20], and degree d of the polynomial kernel over the range {1, 2}. Since theoretical literature on domain characteristics as it relates to optimal parameter selection is not yet developed, the ranges of costs and degrees for optimization were chosen based on previous empirical studies [1, 18]. Different combinations of costs and degrees were exhaustively evaluated by cross-validation.

Clinical Query Filters

The CQ filters are Boolean queries optimized separately for sensitivity, specificity, and accuracy [10]. We applied the exact queries for optimized sensitivity and specificity cited in Pubmed and recently updated with a year 2000 corpus to the text categorization task [2-4].

Estimating Model Performance

We used 5-fold cross-validation that avoids over-fitting to estimate the performance of the learning algorithms [6]. This choice for n provided sufficient high-quality positive samples for training in each category and provided sufficient article samples for the classifiers to learn the models. The cross-validation procedure first divided the data randomly into 5 non-overlapping subsets of documents where the proportion of positive and negative documents in the full dataset is preserved for each subset. Next, the following was repeated 5 times: we used one subset of documents for testing (the "original testing set") and the remaining four subsets for training (the "original training set") of the classifier. The average performance over 5 original testing sets is reported.

In order to optimize parameters of the SVM algorithms, we used another "nested" loop of cross-validation by further splitting each of the 5 original training sets into smaller training sets and validation sets. For each combination of learner parameters, we obtained cross-validation performance and selected the best performing parameters inside this inner loop of cross-validation. We next built a model with the best parameters on the original training set and applied this model to the original testing set. Details about the "nested cross-validation" procedure can be found in [7, 21]. Notice that the final performance estimate obtained by this procedure will be unbiased because each original testing set is used only once to estimate performance of a single model that was built by using training data exclusively.

Applying Filters to Prospective Corpora

We built final machine learning filter models in each category using the 1998-1999 and 1998-2000 corpora and then applied both the final machine learning filter models and the CQ filters to the prospective 2005 corpus. We built the final machine learning filter models by selecting best performing parameters (i.e. cost and degree) and applying these parameters to build final models in each category using all the data. Best parameters were selected by first, dividing the data into 5 non-overlapping subsets preserving positive/ negative proportions. For each set of parameters, we estimated performance using cross-validation over the 5 folds. Average performance across all folds with each set of parameters was recorded. We selected the parameters that built the best performing filter model, and used these parameters to build a final machine learning filter model for each category using all the data.

Comparing CQ Filters to Learning Models

We compared the sensitivity and specificity of the machine learning filter models with the sensitivity and specificity of the respective optimized Boolean CQ filter [10]. The CQ filters return articles with the query terms present, whereas the learning algorithms return a score. To make the comparison, in each fold, we fixed the sensitivity value returned by the sensitivity-optimized CQ filter and varied the threshold for the scored articles until the sensitivity was matched. We report the fixed sensitivity, corresponding specificity, and precision. The same procedure was run for the specificity returned by the optimized specificity CQ filter.

Results

Area under the curve analysis

We built machine learning filter models for treatment, etiology, prognosis, and diagnosis categories using the 1998-1999 and 1998-2000 corpora. In Table IV-1, we report the cross-validation area under the ROC curve for the 1998-1999 and 1998-2000 built machine learning filter models, and area under the ROC curve performance when the machine learning filter models were applied to the entire 2005 corpora in the 4 categories.

Table IV-1 - Top row is cross-validation estimated area under the curve for optimal 1998-1999 and 1998-2000 models. Bottom row is area under the curve for the optimal models applied to 2005 corpora (*no cross-validation applied*). Treat – treatment, Diag – diagnosis, Prog-prognosis, Etio – etiology. ± - is the range of AUC estimates across the 5 folds.

	Treat	**Diag**	**Prog**	**Etio**
X-Val AUC	0.97± .02	0.99 ± .02	0.95± .02	0.95 ± .01
2005 AUC	0.95	0.97	0.94	0.94

The optimal machine learning filter models built using the 1998-1999 and 1998-2000 corpora and applied to the 2005 corpora had performances within the range of estimates of each fold in each cross-validation set. The optimal machine learning filter models were able to discriminate high quality articles from other non-high-quality articles in the 2005 corpora.

Comparison to CQ filters

We applied the CQ filters of Pubmed to the entire 2005 corpora and reported their corresponding sensitivity and specificities in Table IV-2. In all 4 categories, the CQ filters performed well. The support vector machine outperforms the CQ filters in sensitivity, specificity, and precision at fixed sensitivity and specificity levels.

The specificity and sensitivity optimized prognosis CQ filters and specificity optimized etiology CQ filters have lower sensitivity and specificity than previously reported results. The sensitivity optimized prognosis CQ filter (90.0% as reported in [3] vs. 80.0% in the current study), specificity optimized prognosis CQ filter (94.1% as reported in [3] vs. 76.8% in the current study), and the specificity optimized etiology CQ filter (94.9% as reported in [2] vs. 83.9% in the current study) do not perform as expected. Further investigation is necessary to determine the cause of this performance discrepancy and possible solutions.

Table IV-2 – Optimized Support Vector Machine (SVM) compared to Clinical Query Filters fixed at optimal sensitivity and specificity. All values are calculated using the entire 2005 corpora.

Category	Optimized For	Method	Sensitivity	Specificity	Precision
Treatment	Sensitivity	Query Filters	**0.980**	0.710	0.147
		SVM		0.888	0.305
	Specificity	Query Filters	0.803	**0.913**	0.318
		SVM	0.948		0.349
Etiology	Sensitivity	Query Filters	**0.979**	0.435	0.010
		SVM		0.753	0.024
	Specificity	Query Filters	0.681	**0.839**	0.025
		SVM	0.936		0.035
Diagnosis	Sensitivity	Query Filters	**0.956**	0.682	0.01
		SVM		0.884	0.02
	Specificity	Query Filters	0.652	**0.972**	0.07
		SVM	0.821		0.08
Prognosis	Sensitivity	Query Filters	**0.800**	0.707	0.011
		SVM		0.874	0.024
	Specificity	Query Filters	0.800	**0.768**	0.013
		SVM	1.00		0.017

Discussion

These experiments addressed a pertinent and important question for using a filter to identify articles in a corpus. If we built machine learning or apply semi-manually constructed Boolean-based CQ filters using a corpus from a different time period, can we reliably apply these filters to current corpora and identify the high quality articles.

Our results showed that we can identify articles in this 2005 corpus using CQ filters or machine learning filter models. The optimized machine learning filter models built with the 1998-1999 and 1998-2000 corpora from [1] do generalize as estimated by the cross-validation procedure and were able to identify high quality articles accurately in a 2005 corpora as measured by area under the curve. The CQ filters of Pubmed were also able to identify high quality articles. As anticipated by [1], the optimized machine learning filter models generalize well and had superior ability over the optimized CQ filters to identify quality articles in the 2005 corpus.

These results also validate the optimization methods used to build the machine learning filter models and the consistent editorial policies of the ACP Journal Club. The ability of the 1998-1999 and 1998-2000 corpora based machine learning filter models to identify high quality articles in the 2005 corpus imply that the procedure to optimize the machine learning filter model (through cross-validation) is valid and creates robust models and model performance estimates.

Furthermore, the ACP Journal Club is a consistent, stable gold standard. The 1998-1999 and 1998-2000 based corpora machine learning filter models discriminatory power to identify high quality articles succeeds due to consistent article selection in the original and prospective corpora. The machine learning filter models prediction of high quality articles in the 2005 corpora imply that the methodologic criteria for high quality articles has not changed over time, and we may reliably apply these machine learning filter models in current years.

The true purpose of any filter is to identify high quality articles in later corpora. This paper is a step to validating filters for medical information retrieval. Coupled with our previous work [1], we are establishing a foundation for usage of these filters.

In current work, we are systematically evaluating these filters in answering “real-life” clinical questions. As a first step, we have built a proof of concept system at www.ebmsearch.org. How well these filters can assist expert reviewers and their generalization to other categories and domains are open questions that we have experiments underway to answer.

Acknowledgements

The first author acknowledges support from NLM grant LM007948-02. The second author acknowledges support from grant LM007948-01.

References

1. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text Categorization Models for High Quality Article Retrieval in Internal Medicine. J Amer Med Inform Assoc. 2005;12(2):207-216.

2. Wilczynski N, Haynes B. Developing Optimal Search Strategies for Detecting Clinically Sound Causation Studies in MEDLINE. In: Proc AMIA Symposium; 2003; Washington DC; p. 719-23.

3. Wilczynski N, Haynes B. Optimal Search Strategies for Detecting Clinically Sound Prognostic Studies in EMBASE. J Amer Med Inform Assoc. Jul/Aug 2005;12(4):481-485.

4. Haynes B, Wilczynski N. Optimal Search Strategies for retrieving scientifically strong studies of diagnosis from MEDLINE: an analytical survery. BMJ 2004.

5. Wilczynski N, Haynes B. Robustness of Empirical Search Strategies for Clinical Content. In: AMIA; 2002.

6. Weiss S, Kulikowski CA. Computer Systems that Learn. San Mateo, CA: USA M. Kauffman; 1991.

7. Scheffer T. Error estimation and model selection. Technischen Universit at Berlin; 1999.

8. Kearns M, Umesh V. An Introduction to Computational Learning Theory: MIT Press; 1994.

9. Aliferis CF, Statnikov A, Tsamardinos I. Challenges in the Analysis of Mass-Throughput Data. Cancer Informatics. To appear 2006.

10.(Accessed: 03-13-2006), http://www.ncbi.nlm. nih.gov/entrez/query/static/clinical.html.

11.ACP_Journal. Purpose and Procedure. ACP Journal 1999;131(1):A-15 - A-16.

12.PubMed. (Accessed: 3-06-2006), http://www.ncbi.nlm.nih.gov/PubMed/.

13.Salton G, Buckley C. Term weighting approaches in automatic retrieval. Information Processing and Management 1988;24(5):513-523.

14.Porter MF. An algorithm for suffix stripping. Program 1980;14(3):130-137.

15.MEDLINE Stopwords. (Accessed: 3-13-2006), http://biolib.princeton.edu/instruct/MedSW.html.

16.Leopold E, Kindermann J. Text Categorization with Support Vector Machines. How to Represent Texts In Input Space? Machine Learning 2002;46:423-444.

17.Vapnik V. Statistical Learning Theory. New York: Wiley; 1998.

18.Joachims T. Text Categorization With SVMs: Learning With Many Relevant Features. In: Proceedings of the 10th European Conference On Machine Learning; 1998: Springer-Velag.

19.LIBSVM: a library for support vector machines. (Accessed: 3-13-2006), http://www.csie.ntu.edu.tw/~cjlin/libsvm.

20.Morik K, Brockhausen P, Joachims T. Combining statistical learning with a knowledge-based approach - A case study in intensive care monitoring. In: Proc. 16th Int'l Conf. on Machine Learning (ICML-99); 1999.

21.Dudoit S, Van Der Laan MJ. Asymptotics of cross-validated risk estimation in model selection and performance assessment. Working Paper: U.C. Berkeley Division of Biostatistics; 2003 February 5. Report No.: 126.

Extending Text Categorization Filters in Medicine

Abstract

In this study, we evaluated machine learning filter models to identify high quality articles in areas outside of internal medicine and format, purpose, and rigor content categories. In previous studies, we concluded that machine learning filter models identified high quality articles in internal medicine in the treatment, prognosis, diagnosis, and etiology content categories with high area under the receiver operating curve (AUC). In this study, we used a gold standard encompassing 49,028 articles in 161 journals in areas including pediatrics, psychology, and surgery. We built machine learning filter models in 18 content categories and evaluated their discriminatory performance to identify labeled articles using area under the receiver operating curve analysis. We also compared 5 rigor filter models to Pubmed's Clinical Query Filters. The machine learning filter models identified high quality articles with AUC of greater than 0.92 in all 18 content categories. The machine learning filter models showed comparable performance in treatment and superior performance in etiology, prognosis, diagnosis, and clinical prediction guide when compared to Pubmed's Clinical Query Filters at fixed sensitivity and fixed specificity. An implementation of the models is available at www.ebmsearch.org. Machine learning filter models effectively identify high quality articles in content categories and areas outside of internal medicine.

Introduction

Every publication and talk regarding the biomedical literature mentions its current volume, exponential growth, and the growing challenge for health professionals and medical librarians to identify high quality articles applied to evidence based care. Enrico Coiera [1] postulates an impending “information famine” based on Malthus’ law, where, rather than human population needs outgrowing their food sources, the information glut outgrows humans limited ability to find and assimilate high quality information. As sound as Malthus’ logic seemed to be, his predictions of widespread famine did not come to pass. Malthus did not foresee the vast advances in agricultural technology that would feed the world’s population. In a similar vein, we propose that advanced search technology may provide a solution to the current information glut.

If we can increase the *accuracy* of information search technology at a greater rate than information grows, we may avoid the forthcoming “information famine.” As a first step to increasing the accuracy of search technology, we used automated mechanisms to filter the medical literature for identifying content-specific, high-quality medical articles in internal medicine [2].

Background

Researchers proposed and implemented several methods to automatically or semi-automatically filter the medical literature to identify high quality articles. The problem was defined as a classification problem in identifying high-quality content specific articles or not.

Haynes and colleagues created Boolean queries to identify high quality articles. They used a manually built gold standard of 49,028 labeled articles to create Boolean queries that identify clinically relevant articles in the categories of prognosis, diagnosis, etiology, treatment, and clinical prediction guide. Evaluating articles in 161 journals in 2000, six research assistants labeled high quality articles by constructing a gold standard according to content and methodological criteria [3]. The content areas included etiology, prognosis, diagnosis, and treatment, and the methodological criteria were similar to the criteria currently used by the ACP Journal Club [4]. The authors selected terms that would most likely return high quality articles in these content categories based on interviews with expert librarians and clinicians. Valid MeSH terms, publication types, and wildcarded word roots (i.e. random* matching *randomize* and *randomly*) in the title and abstract were collected. Using the above gold standard and the selected terms, they ran an exhaustive search of all disjunctive Boolean set term models of 4 to 5 terms, and evaluated each disjunctive set on an independent document set according to sensitivity, specificity, and precision of returning high quality articles. The optimal Boolean sets were shown to have high sensitivity, specificity, and precision and are currently featured in the clinical queries link in PubMed [5]. This method required interviewing to select terms, a gold standard constructed by an ad-hoc review panel of expert clinicians, reliance on NLM assigned terms, and search of term disjunctions that grows exponentially with the number of search terms.

Other researchers have applied a similar methodology to developing sets of search terms for controlled trials, systematic reviews, and diagnostic articles [6-11].

Another approach to classification is to use citation measures to identify high quality articles. Citation measures capture directly an article's impact and may serve as a proxy for methodological quality. Bernstam and colleagues tested this hypothesis. They proposed raw citation count and the PageRank algorithm as measures of quality for a Society of Surgical Oncology gold standard [12]. They showed that raw citation count identified articles in the gold standard better than PageRank or automated filters designed for another task [13]. In further research, we showed that machine learning filter models designed specifically for the SSOAB gold standard outperformed citation count and PageRank in identifying articles for this specific gold standard. We also extended the analysis to an ACP Journal club gold standard in treatment and etiology with similar findings [14]. It is likely that citation count is, at best, a moderate predictor of medical literature quality.

In more recent years, Google Scholar arose as a means to use citation counts to measure impact. Though the exact algorithms used by Google are proprietary, the basic tenet involves ranking articles by their citation counts [15]. We postulate that the moderate correlation between citation count and quality should extend to the web as well.

The moderate correlation between citation count and quality is based on the idea that not all citations, whether bibliographical or web based, are necessarily endorsements of the article or page. An article may cite another article for a variety of reasons: authors may cite articles to acknowledge prior work, identify methodology, provide background reading, correct or criticize, substantiate claims, alert readers to forthcoming work, authenticate data, identify original publication of a term or concept, disclaim work of others, or dispute priority claims [16]. The lack of an unambiguous connection between citation, context of use, manner of use, and/or endorsement prevents citation count from being a single effective measure of inclusion in an “importance” bibliography. More generally stated, the conceivable reasons for citation are so numerous that it is unrealistic to believe that citation conveys just one semantic interpretation. Other research with medical datasets seems to support this weak relationship between citation counts and quality [17-20].

A promising approach to classification is in the use of text categorization techniques to identify high quality articles. We applied advanced pattern recognition techniques to identify high quality articles in internal medicine [2]. We constructed a high quality corpus with labeled high quality articles in etiology, prognosis, diagnosis, and treatment. We used 10 fold cross-validation techniques to estimate performance and were able to identify high quality articles with high discriminatory performance as measured by area under the curve (AUC) and with better performance than corresponding Boolean clinical query filters built for the same task.

In later work, we validated that models built from an earlier corpus identified high quality articles in a later corpus [21]. Furthermore, we showed that the models perform better than general citation counts such as PageRank and raw bibliographic citation count in identifying high quality articles [14]. We showed that in treatment related clinical questions, the models perform on par if not better in identifying expert librarian selected articles that answered the clinical questions [22].

Though the previous studies showed that the validated models performed better than general citations metrics and identified articles that could answer treatment related clinical questions, the experiments had several limitations. The models were only proven to work in internal medicine. The machine learning filter models were compared to Pubmed's Clinical Query Filters which were built with a different corpus than the validated filter models. Finally, we only evaluated these filters for four semantic categories.

Hypothesis

In this study, we address the shortcomings of the previous work and address 5 hypotheses. First, we hypothesize that the models work in areas outside of internal medicine. The models work in other semantic categories including clinical prediction guide, costs, and economics. The models generalize to purpose and format categories. The models work better than the Clinical Query Filters when directly compared. Finally, these filters can be implemented in a practical system.

Methods

Gold Standard Construction

We used a rigorous gold standard developed by the Hedges group [3]. Haynes and colleagues trained 6 research assistants to rate articles from 161 journals for the publishing year 2000. They rated each article by purpose and quality in the content areas of treatment, diagnosis, prognosis, etiology, economics, clinical prediction guide, qualitative, and review. Furthermore, articles were labeled by format in the areas of clinically relevant original studies, review articles, general papers, or case reports.

The research assistants were rigorously calibrated and inter rater agreement for methodologic criteria exceeded 80% beyond chance. Some methodologic criteria for ranking are given in Table IV-3.

Table IV-3 – Labeling Criteria

Purpose/ Rigor

Class	Category	Criteria
Format	Original	Any full text article in which the investigators report first-hand observations.
	Review	Any full text article that is bannered ‘review, overview, or meta-analysis’ in the title or in a section heading, or it is indicated in the text of the article that the intention was to review, summarize, highlight, etc. the literature on a particular topic.
	General and Miscellaneous Articles	A general or philosophical discussion of a topic without original observation and without a statement that the purpose was to review or appraise a body of knowledge. This could include news items, unbannered editorials, bannered and unbannered conference reports, position and opinion papers, musings, psychosocial observations, and decision analysis that cannot be classified as an original study or review.
	Case Report	Is an original study or report that presents only individualized data. The data are not combined in any way, and often involves less than 10 subjects. If the article is a CR do not fill out as an original study. If the article also states that it is a review of the literature, fill out a second line for review article.
Purpose	Etiology	Content pertains directly to determining if there is an association (causal link) between an exposure and a disease or condition (examples of a condition are low birth weight, large [or small] for gestational age, preterm birth, miscarriage, abortion, cesarean section, pregnancy, or death). The

		question that is being asked is “What causes people to get a disease or condition?”
	Prognosis	Content pertains directly to the prediction of the clinical course or the natural history of a disease or condition (examples of a condition are low birth weight, large [or small] for gestational age, preterm birth, or pregnancy) with the disease or condition existing at the beginning of the study.
	Diagnosis	Content pertains directly to using a tool to arrive at a diagnosis of a disease or condition. Screening to make a diagnosis is included here.
	Treatment	Content pertains directly to therapy (including adverse effects studies), prevention, rehabilitation, quality improvement, or continuing medical education. For a study to be classified as therapy (which includes prevention, continuing medical education and quality improvement) the investigators must intervene – there has to be an intervention that can be manipulated.
	Costs	Content pertains directly to the costs or financing or economics of a health care issue.
	Economics	Content pertains directly to the economics of a health care issue. The economic question addressed must be based on comparison of alternatives, i.e., comparison of the costs and effects of at least 2 different forms of intervention or service provision. Thus, ‘costing’ or ‘financing’ of a single health service, even if for a variety of conditions, does not constitute an economic study; an economic study would compare 2 (or more) different ways of providing the same service, and would include at least intermediate (e.g., BP) or more advanced (e,g., stroke) outcomes. Economics studies are also Costs studies and should be co-classified there.

	Clinical Prediction Guide	Content pertains directly to the prediction of some aspect of a disease or condition; the authors must indicate that the purpose of the study is to develop or validate a rule, guide, index, equation, scale, score or model to predict a diagnosis, prognosis, risk (ET), therapeutic response, therapeutic drug levels or clinical outcome. For everything except diagnosis the patients must be followed over time.
	Qualitative	Content of study contains the following qualities: The content relates to how people feel or experience certain situations, specifically those situations that relate to health care in humans. Collection methods are appropriate for qualitative data. Analyses are appropriate for qualitative data.
Rigor	Treatment	Random allocation of participants to comparison groups; Outcome assessment of at least 80% of those entering the investigation; Analysis consistent with study design.
	Etiology	Observations concerned with the relationship between exposures and putative clinical outcomes; Data collection is prospective; Clearly identified comparison group(s); Blinding of observers of outcome to exposure.
	Diagnosis	Inclusion of a spectrum of participants; Objective diagnostic ("gold") standard OR current clinical standard for diagnosis; Participants received both the new test and some form of the diagnostic standard; Interpretation of diagnostic standard

		without knowledge of test result and visa versa; Analysis consistent with study design.
	Prognosis	Inception cohort of individuals all initially free of the outcome of interest; Follow-up of at least 80% of patients until the occurrence of a major study end point or to the end of the study; Analysis consistent with study design.
	Clinical Prediction Guide	Guide is generated in one or more sets of real patients (training set); Guide is validated in an independent set of real patients (test set).
	Qualitative	Methodologic rigor is not evaluated for qualitative studies.
	Cost	Methodologic rigor is not evaluated for costs studies. Economics studies are a subset of costs studies and should be indicated so; economics studies should then be further evaluated under Economics (see next).
	Economics	Question is a comparison of alternatives; Alternative services or activities compared on outcomes produced (effectiveness) and resources consumed (costs); Evidence of effectiveness must be from a study of real patients that meets the above-noted criteria for diagnosis, treatment, quality improvement, or a systematic review article; Effectiveness and cost estimates based on individual patient data (micro-economics); Results presented in terms of the incremental or additional costs and

		outcomes of one intervention over another; Sensitivity analysis if there is uncertainty.

The selected journals encompass areas outside of internal medicine including pediatrics and the surgical specialties. These additional journals include the “American Journal of Surgery,” “Annals of Surgery,” “Archives of Surgery,” “Clinical Pediatrics,” “Journal of Pediatrics,” “Obstetrics and Gynecology,” and others. Table IV-4 shows a random selection of 79 out of the 161 rated journals. A more comprehensive list is available from the authors.

Table IV-4 – 79 out of 161 Randomly Selected Journals Reviewed for Hedges Corpora.

A more comprehensive list is available from the authors.

- AJR American Journal of Roentgenology
- Acta Orthopaedica Scandinavica
- American Journal of Cardiology
- American Journal of Gastroenterology
- American Journal of Medicine
- American Journal of Obstetrics & Gynecology
- American Journal of Public Health
- American Journal of Surgery
- Annals of Emergency Medicine
- Annals of Medicine
- Annals of the Rheumatic Diseases
- Archives of Disease in Childhood
- Archives of Family Medicine
- Archives of Medical Research
- Archives of Surgery
- Arthritis & Rheumatism
- Australian & New Zealand Journal of Psychiatry
- Birth
- Canadian Journal of Gastroenterology
- Canadian Journal of Psychiatry Revue
- Canadienne de Psychiatrie
- Cancer
- Chest
- Clinical & Investigative Medicine Medecine
- Clinique et Experimentale
- Clinical Psychology Review
- Cochrane database of systematic reviews computer file
- Critical Care Medicine
- Development & Psychopathology
- Diabetes Care
- Diabetic Medicine
- Family Planning Perspectives
- Family Practice
- Gastroenterology
- Gut
- Health Education & Behavior
- Heart & Lung
- Injury
- International Journal of Geriatric Psychiatry
- JAMA
- Journal of Abnormal Child Psychology
- Journal of Arthroplasty
- Journal of Autism & Developmental Disorders
- Journal of Child & Adolescent Psychopharmacology
- Journal of Clinical & Experimental Neuropsychology
- Journal of Clinical Child Psychology
- Journal of Clinical Epidemiology
- Journal of Clinical Nursing
- Journal of Clinical Psychopharmacology
- Journal of Consulting & Clinical Psychology
- Journal of Family Practice
- Journal of Infectious Diseases
- Journal of Internal Medicine
- Journal of Manipulative & Physiological Therapeutics
- Journal of Neuropsychiatry & Clinical Neurosciences
- Journal of Orthopaedic Research
- Journal of Pediatrics
- Journal of Psychosomatic Research
- Journal of Rheumatology
- Journal of Trauma Injury Infection & Critical Care
- Journal of Vascular Surgery
- Journal of the American College of Cardiology
- Journal of the American Geriatrics Society
- Journal of the American Medical Informatics Association
- Lancet
- Medical Care
- Medical Journal of Australia
- Midwifery
- Neurology
- Nursing Research
- Patient Education & Counseling
- Plastic & Reconstructive Surgery
- Psychiatric Services
- Psychological Medicine
- Psychology & Aging
- Psychosomatic Medicine
- Public Health Nursing
- Radiology
- Social Science & Medicine
- Stroke
- Thorax

Article Preparation

The conversion of documents to a format suitable for the machine learning algorithm followed the procedures in [2]. The articles in the ACPJ selected journals were cross-referenced in PubMed, and the title, abstract, journal, publication type, and MeSH terms were extracted. We created two representations for each document: one for the machine learning algorithm, and one for the clinical query (CQ) filters.

For the machine learning algorithm, we represented each document as a set of terms for the learning algorithms [23]. We stemmed each term [24], removed "stopword" terms [25], and removed any terms occurring in fewer than 5 documents. Very infrequent terms are difficult to assess statistically and may affect negatively the generalization of the classification models. Terms were further encoded as weighted features using a log frequency with redundancy scheme [26].

For the CQ filters, we represented each document as a set of terms. Words were not stemmed, but "stopwords" and infrequent terms (occurring in < 5 documents) were removed.

Statistical and Machine Learning Methods

Support Vector Machines (SVMs)

In our experiments, we employed Support Vector Machine (SVM) classification algorithms. The SVM's calculate maximal margin hyperplane(s) separating two or more classes of the data. To accomplish this, the data are mapped to a higher dimensional space by means of a kernel function, where a separating hyperplane is found by solving a constrained quadratic optimization problem [27]. SVMs have had superior text classification performance compared to other methods [2, 28], and this motivated our use of them. We used an SVM classifier implemented in libSVM v2.83 [29] with a polynomial kernel. We optimized the SVM penalty parameter C over the range {0.1, 1, 2} with imbalanced costs applied to each class proportional to the priors in the data [30], and degree d of the polynomial kernel over the range {1, 2}. Since theoretical literature on domain characteristics as it relates to optimal parameter selection is not yet developed, the ranges of costs and degrees for optimization were chosen based on previous empirical studies [2, 28]. Different combinations of costs and degrees were exhaustively evaluated by cross-validation.

Clinical Query Filters

The CQ filters are Boolean queries optimized separately for sensitivity, specificity, and accuracy [31]. We applied the exact queries built with this gold standard optimized for sensitivity and specificity cited in Pubmed [6, 32, 33]. Queries in treatment, diagnosis, prognosis, etiology, and clinical prediction guide were compared to the corresponding models.

Estimating Model Performance

We used cross-validation to estimate the performance of the learning algorithms [34]. This choice for n provided sufficient high-quality positive samples for training in each category and provided sufficient article samples for the classifiers to learn the models. The cross-validation procedure first divided the data randomly into 5 non-overlapping subsets of documents where the proportion of positive and negative documents in the full dataset is preserved for each subset. Next, the following was repeated 5 times: we used one subset of documents for testing (the "original testing set") and the remaining four subsets for training (the "original training set") of the classifier. The average performance over 5 original testing sets is reported.

In order to optimize parameters of the SVM algorithms, we used another "nested" loop of cross-validation by further splitting each of the 5 original training sets into smaller training sets and validation sets. For each combination of learner parameters, we obtained cross-validation performance and selected the best performing parameters inside this inner loop of cross-validation. We next built a model with the best parameters on the original training set and applied this model to the original testing set. Details about the "nested cross-validation" procedure can be found in [35, 36]. Notice that the final performance estimate obtained by this procedure will be unbiased because each original testing set is used only once to estimate performance of a single model that was built by using training data exclusively.

Comparing CQ Filters to Learning Models

We compared the sensitivity and specificity of the machine learning filter models with the sensitivity and specificity of the respective optimized Boolean CQ filter [31]. The CQ filters return articles with the query terms present, whereas the learning algorithms return a score. To make the comparison, in each fold, we fixed the sensitivity value returned by the sensitivity-optimized CQ filter and varied the threshold for the scored articles until the sensitivity was matched. We report the fixed sensitivity, corresponding specificity, and precision. The same procedure was run for the specificity returned by the optimized specificity CQ filter.

Building final models

We built final machine learning filter models in each category using the Hedges corpora and applied the final machine learning filter models to 13 million documents in the MEDLINE article collection. We built the final machine learning filter models by selecting best performing parameters (i.e. cost and degree) and applying these parameters to build final models in each category using *all* the data. Best parameters were selected by first, dividing the data into 5 non-overlapping subsets preserving positive/ negative proportions. For each set of parameters, we estimated performance using cross-validation over the 5 folds. Average performance across all folds with each set of parameters was recorded. We selected the parameters that built the best performing filter model and used these parameters to build a final machine learning filter model for each category using all the data [37].

Building a system

We implement a system called EBMSearch located at http://www.ebmsearch.org. EBMSearch applied these models to articles published between 2000 and 2006 in MEDLINE. Users select format and purpose and rigor categories of cost, economics, clinical prediction guide, diagnosis, prognosis, etiology, and treatment, and time frames of 1 year, 2 years, and 5 years for ranking the returned results. The resulting list of articles is ranked by SVM model output score.

The EBMSearch query mirrors the functionality of a Pubmed query. EBMSearch supports field descriptions and tagged search by text word and location (i.e. title, abstract, etc), author, journal title, date, phrase, gender, language, age group, or human or animal studies [38]. EBMSearch also supports automatic term mapping for untagged terms to match, in order, terms in a MeSH translation table, a Journals translation table, Full Author translation table, and an Author index [39]. Any Pubmed query is an acceptable EBMSearch query. In addition, the user specifies a search category (i.e. treatment, diagnosis, prognosis, etiology, clinical prediction guide, and qualitative studies) and a time frame to search (i.e. 1 year, 2 years, 5 years).

EBMSearch sorts the returned articles by score from the category specific SVM model. The SVM model output score is a relative value denoting how statistically similar terms in the article matched terms in the gold standard articles. We defined relevancy as articles similar to the gold standard articles. We generally state that model scores below 0 are *not* relevant, and scores above 0 are relevant. Scores above 1 are more relevant and scores below 1 are less relevant. All articles from the query are sorted and shown from highest to lowest score regardless of relative value. Ranking has proven popular with other major search engines including Pubmed, Yahoo, Google, and MSNSearch.

Results

Area under the curve analysis

We built machine learning filter models for clinical prediction guide, cost, diagnosis, economics, etiology, prognosis, qualitative, and treatment purpose categories. We also built models for the rigor categories of clinical prediction guide, cost, diagnosis, etiology, prognosis, treatment and economics, and the format categories of case reports, original, review, and general miscellaneous. In Table IV-5, we report the 5 fold cross-validation area under the ROC curve with ranges for the machine learning filter models. In all purpose, rigor, and format categories, the machine learning filter models discriminate articles with area under the curves of greater than 0.926 and close ranges across the five folds. Composite ROC curves for each class are shown in Figure IV-1, Figure IV-2, Figure IV-3, and Figure IV-4.

Figure IV-1 - Format Category Receiver Operating Curves

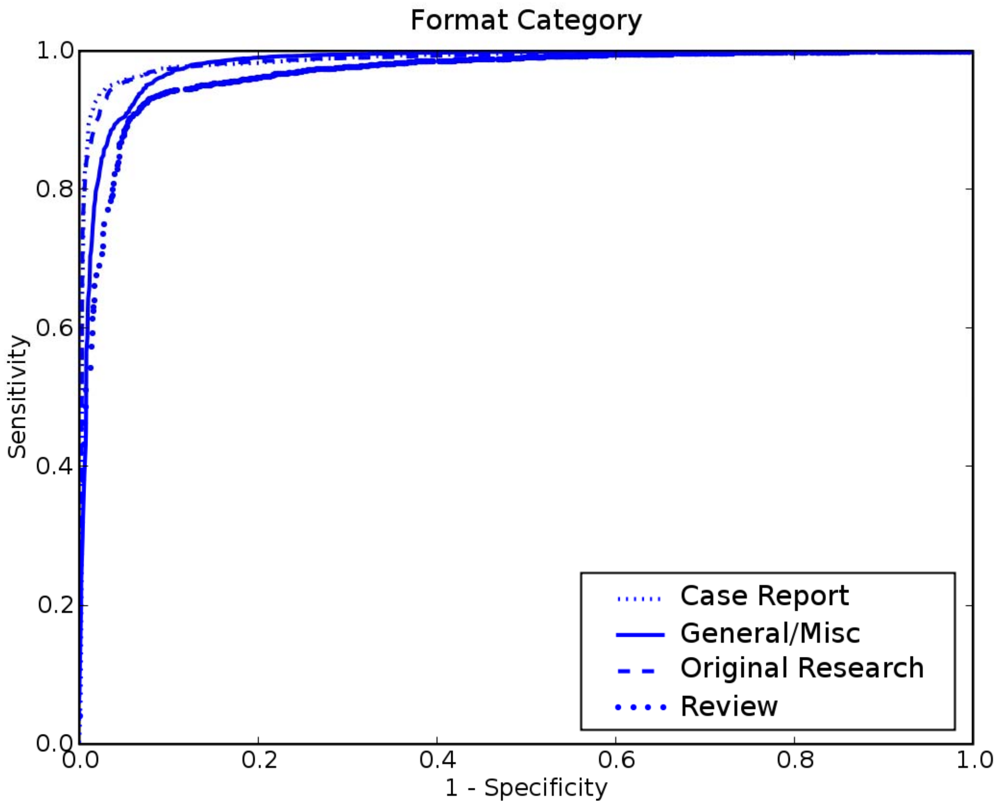


Figure IV-2 - Rigor Category Receiver Operating Curves

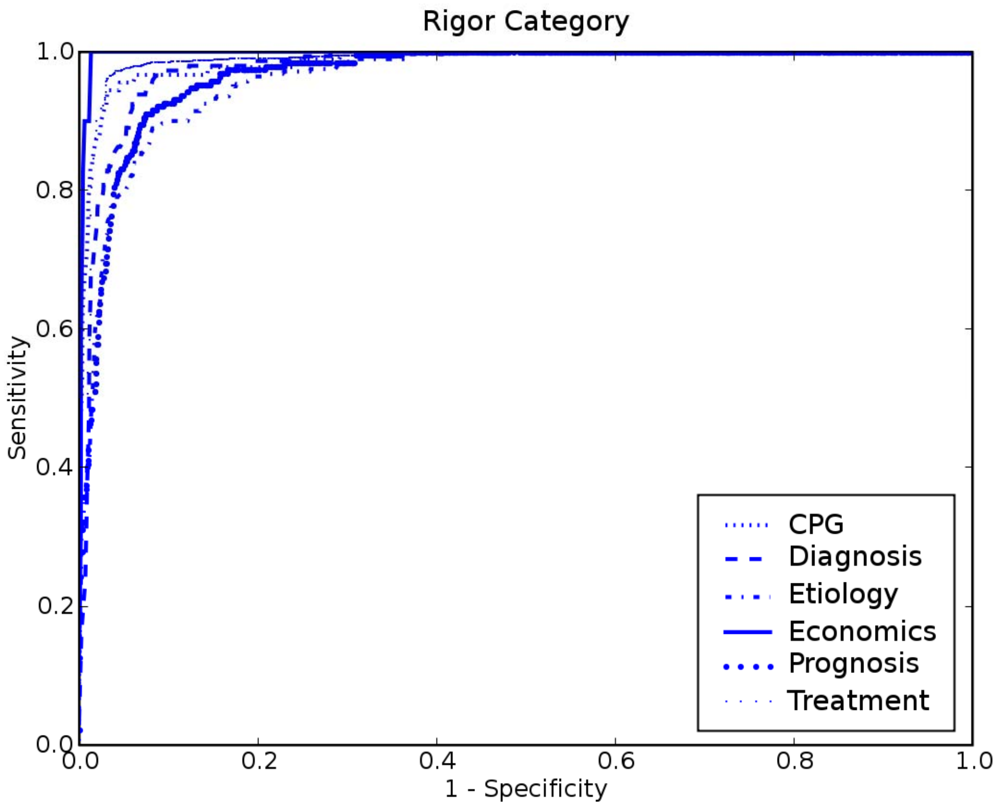


Figure IV-3 - Purpose Category Receiver Operating Curves (Cost, CPG, Diagnosis, Etiology)

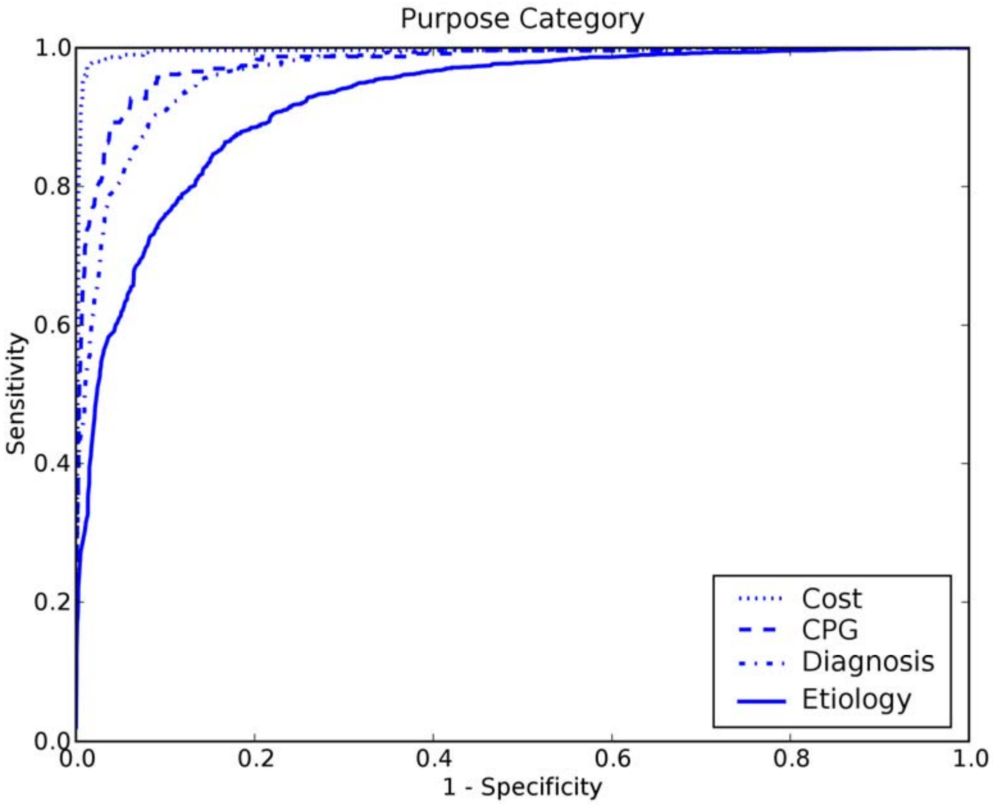


Figure IV-4 - Purpose Category Receiver Operating Curves (Economics, Prognosis, Qualitiative, Treatment)

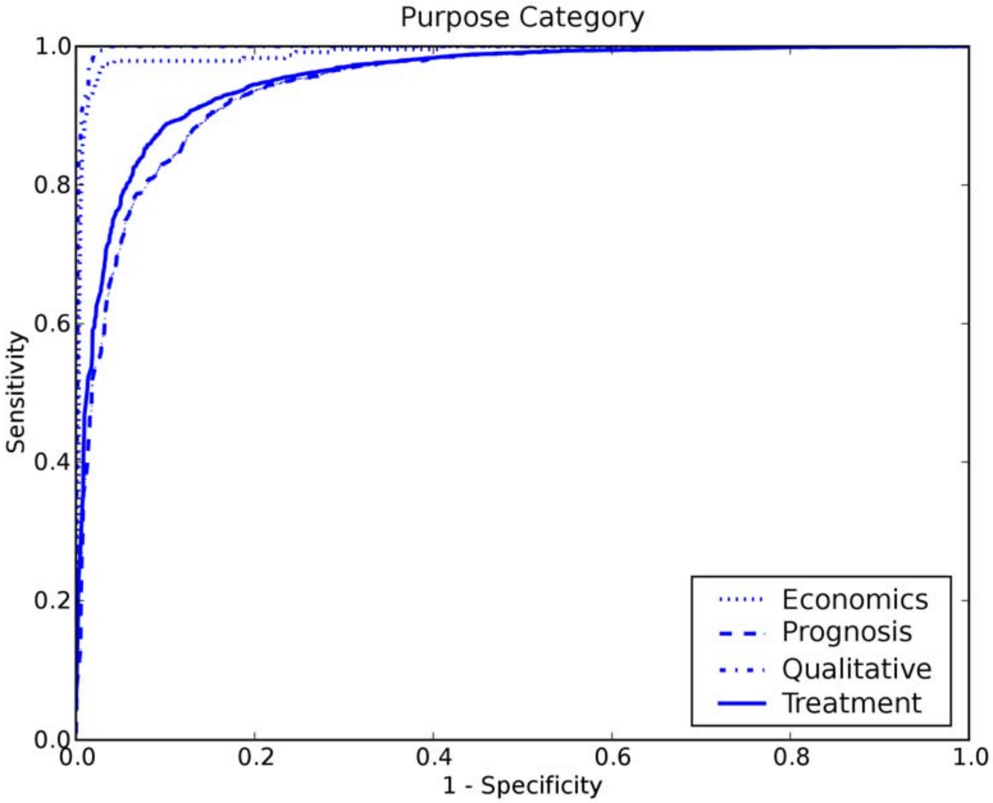


Table IV-5 – AUC for specific categories.

Class	**Category**	**Positives**	**Negatives**	**AUC**
Purpose	Clinical Prediction Guide	232	48796	0.980 (0.966 - 0.989)
Purpose	Cost	300	48728	0.995 (0.987 - 0.998)
Purpose	Diagnosis	1114	47914	0.966 (0.958 - 0.973)
Purpose	Economics	236	48792	0.991 (0.984 - 0.997)
Purpose	Etiology	3018	46010	0.926 (0.920 - 0.931)
Purpose	Prognosis	1642	47386	0.942 (0.939 - 0.945)
Purpose	Qualitative	336	48692	0.997 (0.997 - 0.998)
Purpose	Treatment	8328	40700	0.953 (0.952 - 0.954)
Rigor	Clinical Prediction Guide	91	48937	0.985 (0.974 - 0.995)
Rigor	Diagnosis	147	48881	0.982 (0.972 - 0.992)
Rigor	Etiology	281	48747	0.962 (0.955 - 0.976)
Rigor	Prognosis	190	48838	0.963 (0.954 - 0.970)
Rigor	Treatment	1587	47441	0.988 (0.983 - 0.992)
Rigor	Economics	34	48994	0.997 (0.993 - 0.998)
Format	Case Reports	4591	43461	0.986 (0.985 - 0.989)
Format	Original	25750	22302	0.985 (0.984 - 0.987)
Format	Review	3097	44955	0.970 (0.965 - 0.973)
Format	GM	14747	33305	0.980 (0.980 - 0.982)

Comparison to CQ filters

We applied the CQ filters of Pubmed to the 2000 corpora and reported their corresponding sensitivity and specificities in Table IV-6. In all 5 categories, the CQ filters performed well. Overall, the machine learning filter models outperform the CQ filters in sensitivity, specificity, and precision at fixed sensitivity and specificity levels.

In treatment, at fixed sensitivity and specificity, the machine learning models have similar performance compared to the CQ filters.

In etiology, the machine learning filter models have 0.21 higher specificity at fixed sensitivity than the CQ filters. At fixed specificity in etiology, the models have 0.27 higher sensitivity than the CQ filters.

In diagnosis, the machine learning filter models have 0.19 higher specificity at fixed sensitivity than the CQ filters. At fixed specificity, the filter models have a slight advantage of 0.045 higher sensitivity.

In prognosis, the machine learning filter models have 0.13 higher specificity at fixed sensitivity than the CQ filters. At fixed sensitivity, the filter models have a 0.20 advantage over the CQ filters.

In clinical prediction guides, the machine learning filter models have slight advantage in sensitivity and specificity at both fixed sensitivity and specificity (0.098 higher specificity at fixed sensitivity and 0.052 higher sensitivity at fixed specificity).

The exact splits used to build the Boolean queries were not available to replicate the Boolean query building methods. Though the Boolean queries perform well, it is possible

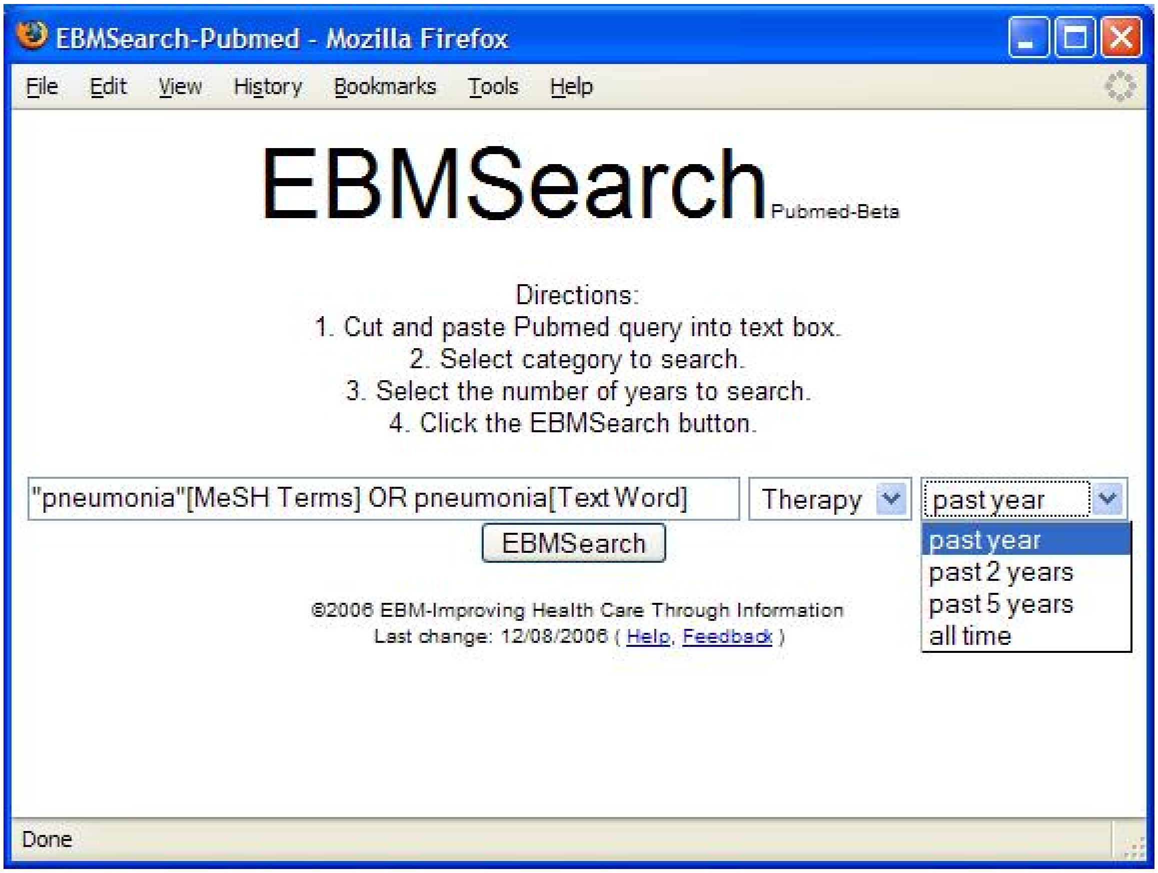
Table IV-6 – Sensitivity/ Specificity Optimized CQF Comparison

Category	Optimized For	Method	Sensitivity	Specificity	Precision
Treatment	Sensitivity	Query Filters	**0.990 (0.978 - 1.000)**	0.786 (0.775 - 0.793)	0.134 (0.127 - 0.138)
		SVM		0.798 (0.594 - 0.963)	0.244 (0.076 - 0.462)
	Specificity	Query Filters	0.950 (0.937 - 0.966)	**0.972 (0.969 - 0.974)**	0.529 (0.509 - 0.547)
		SVM	0.970 (0.968 - 0.972)		0.526 (0.507 - 0.544)
Etiology	Sensitivity	Query Filters	**0.934 (0.911 - 0.946)**	0.631 (0.626 - 0.636)	0.014 (0.014 - 0.015)
		SVM		0.841 (0.824 - 0.877)	0.033 (0.030 - 0.041)
	Specificity	Query Filters	0.565 (0.429 - 0.737)	**0.938 (0.934 - 0.941)**	0.050 (0.038 - 0.067)
		SVM	0.836 (0.821 - 0.857)		0.071 (0.069 - 0.073)
Diagnosis	Sensitivity	Query Filters	**0.990 (0.966 - 1.000)**	0.728 (0.723 - 0.732)	0.011 (0.010 - 0.011)
		SVM		0.918 (0.914 - 0.937)	0.035 (0.033 - 0.043)
	Specificity	Query Filters	0.645 (0.517 - 0.724)	**0.984 (0.983 - 0.984)**	0.106 (0.082 - 0.122)
		SVM	0.690 (0.690 - 0.690)		0.105 (0.102 - 0.109)
Prognosis	Sensitivity	Query Filters	**0.863 (0.763 - 0.921)**	0.798 (0.791 - 0.803)	0.016 (0.014 - 0.018)
		SVM		0.928 (0.880 - 0.965)	0.050 (0.029 - 0.077)
	Specificity	Query Filters	0.516 (0.316 - 0.711)	**0.940 (0.939 - 0.941)**	0.032 (0.020 - 0.043)
		SVM	0.816 (0.816 - 0.816)		0.050 (0.049 - 0.051)
Clinical Prediction Guide	Sensitivity	Query Filters	**0.947 (0.929 – 1.000)**	0.794 (0.787 - 0.799)	0.008 (0.007 - 0.013)
		SVM		0.892 (0.706 – 0.939)	0.024 (0.006 - 0.028)
	Specificity	Query Filters	0.559 (0.412 – 0.714)	**0.993 (0.991 - 0.994)**	0.124 (0.086 – 0.167)
		SVM	0.611 (0.556 – 0.667)		0.126 (0.117 – 0.137)

that the queries are overfit to the data since there is likely overlap between training and testing data.

EBMSearch Implementation

Figure IV-5 shows the front page of the EBMSearch proof of concept system that implements the models. Users input Pubmed formatted queries, a category, and a time period. The result page is shown in



Directions:
1. Cut and paste Pubmed query into text box.
2. Select category to search.
3. Select the number of years to search.
4. Click the EBMSearch button.

"pneumonia"[MeSH Terms] OR pneumonia[Text Word]

Therapy

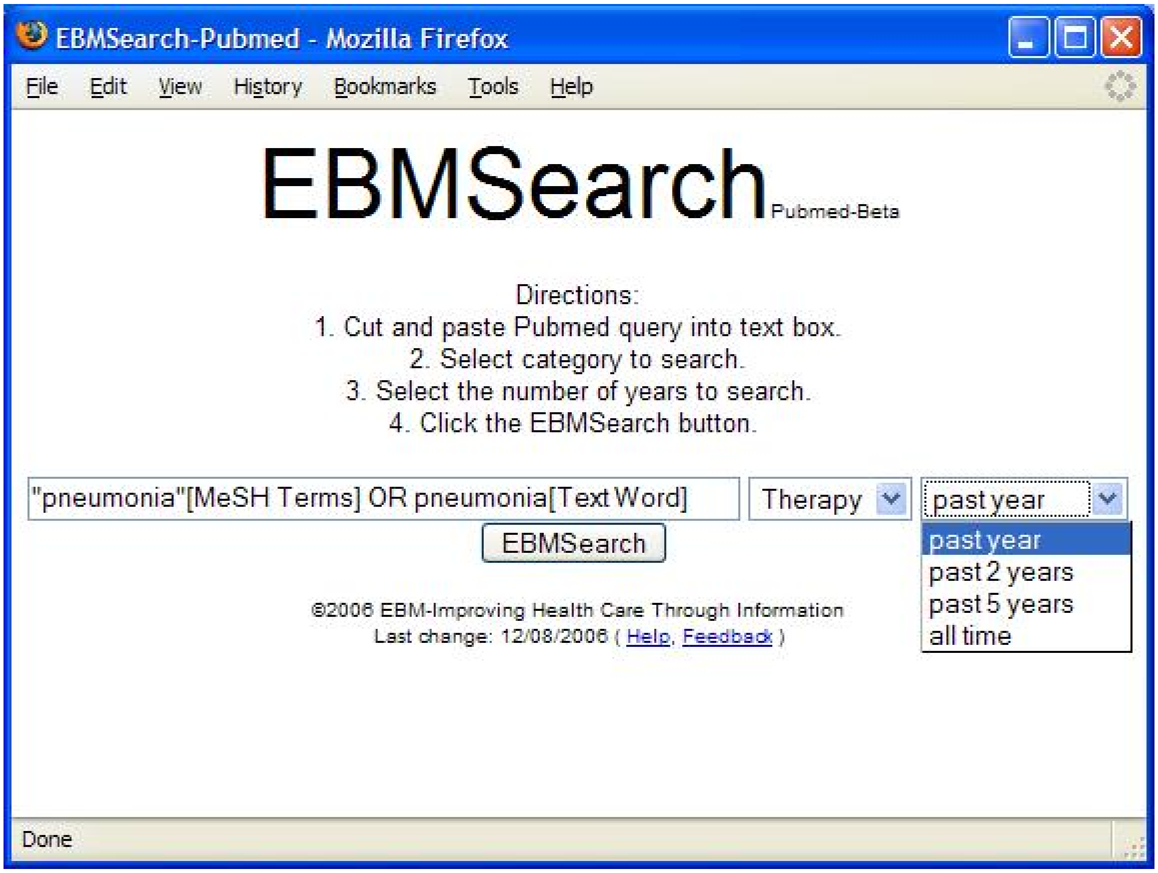
past year

©2006 EBM-Improving Health Care Through Information
Last change: 12/08/2006 (Help, Feedback)

Figure IV-6. Results are rank ordered with sections and query terms highlighted. Clicking on the citation title will send the user to the corresponding Pubmed page.

We refer the reader to [22] for a preliminary evaluation of a subset of these models in answering clinical questions.

Figure IV-5 – Start page for EBMSearch.org



EBMSearch-Pubmed - Mozilla Firefox

EBMSearch Pubmed-Beta

Directions:
1. Cut and paste Pubmed query into text box.
2. Select category to search.
3. Select the number of years to search.
4. Click the EBMSearch button.

"pneumonia"[MeSH Terms] OR pneumonia[Text Word]

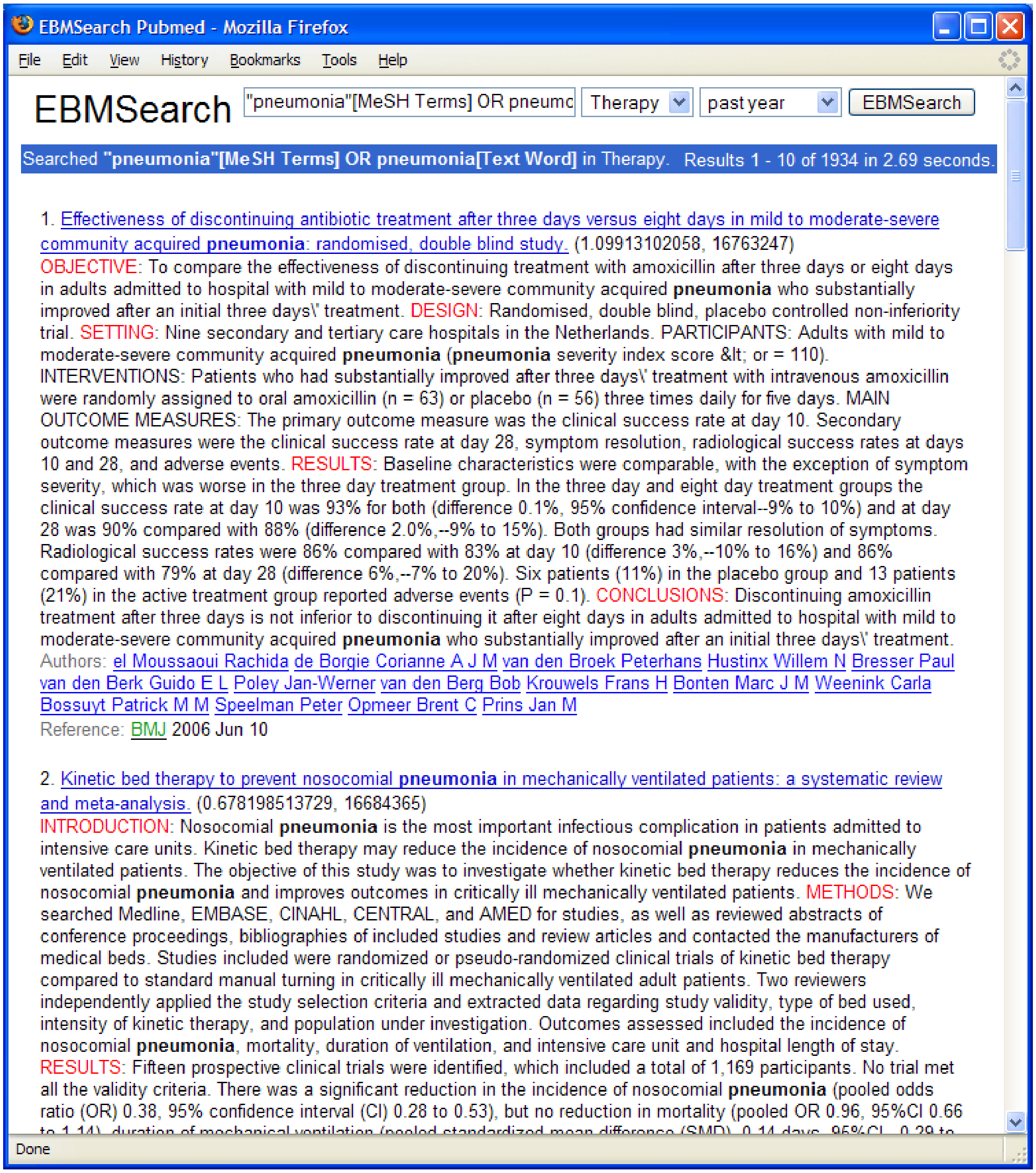
Therapy

past year

©2006 EBM-Improving Health Care Through Information
Last change: 12/08/2006 (Help, Feedback)

Done

Figure IV-6 – Results Page for EBMSearch.org


EBMSearch Pubmed - Mozilla Firefox
File Edit View History Bookmarks Tools Help
EBMSearch
"pneumonia"[MeSH Terms] OR pneumo
Therapy
past year
EBMSearch
Done

Searched **"pneumonia"[MeSH Terms] OR pneumonia[Text Word]** in Therapy. Results 1 - 10 of 1934 in 2.69 seconds.

1. Effectiveness of discontinuing antibiotic treatment after three days versus eight days in mild to moderate-severe community acquired **pneumonia**: randomised, double blind study. (1.09913102058, 16763247)
OBJECTIVE: To compare the effectiveness of discontinuing treatment with amoxicillin after three days or eight days in adults admitted to hospital with mild to moderate-severe community acquired **pneumonia** who substantially improved after an initial three days\' treatment. DESIGN: Randomised, double blind, placebo controlled non-inferiority trial. SETTING: Nine secondary and tertiary care hospitals in the Netherlands. PARTICIPANTS: Adults with mild to moderate-severe community acquired **pneumonia** (**pneumonia** severity index score &lt; or = 110). INTERVENTIONS: Patients who had substantially improved after three days\' treatment with intravenous amoxicillin were randomly assigned to oral amoxicillin (n = 63) or placebo (n = 56) three times daily for five days. MAIN OUTCOME MEASURES: The primary outcome measure was the clinical success rate at day 10. Secondary outcome measures were the clinical success rate at day 28, symptom resolution, radiological success rates at days 10 and 28, and adverse events. RESULTS: Baseline characteristics were comparable, with the exception of symptom severity, which was worse in the three day treatment group. In the three day and eight day treatment groups the clinical success rate at day 10 was 93% for both (difference 0.1%, 95% confidence interval--9% to 10%) and at day 28 was 90% compared with 88% (difference 2.0%,--9% to 15%). Both groups had similar resolution of symptoms. Radiological success rates were 86% compared with 83% at day 10 (difference 3%,--10% to 16%) and 86% compared with 79% at day 28 (difference 6%,--7% to 20%). Six patients (11%) in the placebo group and 13 patients (21%) in the active treatment group reported adverse events (P = 0.1). CONCLUSIONS: Discontinuing amoxicillin treatment after three days is not inferior to discontinuing it after eight days in adults admitted to hospital with mild to moderate-severe community acquired **pneumonia** who substantially improved after an initial three days\' treatment.
Authors: el Moussaoui Rachida de Borgie Corianne A J M van den Broek Peterhans Hustinx Willem N Bresser Paul van den Berk Guido E L Poley Jan-Werner van den Berg Bob Krouwels Frans H Bonten Marc J M Weenink Carla Bossuyt Patrick M M Speelman Peter Opmeer Brent C Prins Jan M
Reference: BMJ 2006 Jun 10

2. Kinetic bed therapy to prevent nosocomial **pneumonia** in mechanically ventilated patients: a systematic review and meta-analysis. (0.678198513729, 16684365)
INTRODUCTION: Nosocomial **pneumonia** is the most important infectious complication in patients admitted to intensive care units. Kinetic bed therapy may reduce the incidence of nosocomial **pneumonia** in mechanically ventilated patients. The objective of this study was to investigate whether kinetic bed therapy reduces the incidence of nosocomial **pneumonia** and improves outcomes in critically ill mechanically ventilated patients. METHODS: We searched Medline, EMBASE, CINAHL, CENTRAL, and AMED for studies, as well as reviewed abstracts of conference proceedings, bibliographies of included studies and review articles and contacted the manufacturers of medical beds. Studies included were randomized or pseudo-randomized clinical trials of kinetic bed therapy compared to standard manual turning in critically ill mechanically ventilated adult patients. Two reviewers independently applied the study selection criteria and extracted data regarding study validity, type of bed used, intensity of kinetic therapy, and population under investigation. Outcomes assessed included the incidence of nosocomial **pneumonia**, mortality, duration of ventilation, and intensive care unit and hospital length of stay. RESULTS: Fifteen prospective clinical trials were identified, which included a total of 1,169 participants. No trial met all the validity criteria. There was a significant reduction in the incidence of nosocomial **pneumonia** (pooled odds ratio (OR) 0.38, 95% confidence interval (CI) 0.28 to 0.53), but no reduction in mortality (pooled OR 0.96, 95%CI 0.66 to 1.14), duration of mechanical ventilation (pooled standardized mean difference (SMD) -0.14 days, 95%CI -0.29 to

Discussion

The machine learning filter models are robust and versatile in areas outside of internal medicine including pediatrics and surgery, other categories including cost, economics, and clinical prediction guide, and other article classes including format and purpose categories. The machine learning models also have equal or better performance than the clinical query filters at fixed sensitivity and specificity.

Comparison to Clinical Query Filters

The comparison between the clinical query filters and the machine learning filter models is more valid compared to previous studies. In previous studies [2], we compared the machine learning filter models built on one gold standard to the clinical query filters built on *another* gold standard. This design may bias the results of the learning algorithm. In this study, the gold standards are identical. In the etiology, diagnosis, and prognosis categories, the machine learning filters identify articles with better sensitivity and specificity at fixed specificity and sensitivity respectively. In the treatment and clinical prediction guide categories, the performance at both fixed sensitivity and specificity are nearly identical.

Clinical query filters were not available in the economics rigor category, or the purpose categories of clinical prediction guide, cost, diagnosis, economics, etiology, prognosis, qualitative, or treatment, or the format categories of case reports, original, review, or general miscellaneous. Thus, we made no comparisons between the machine learning filter models to these other categories.

The results of the clinical query filters may be biased high because of training and test set overlaps. The original train/ test splits to build the clinical query filters were not available. We generated each train/test cross validation split randomly without knowing whether the data used to build the original clinical query filters overlapped with data in each test split. Likely, this design resulted in data overlap between the training and test sets. Even with this bias though, the machine filter models have superior performance in etiology, diagnosis, prognosis, and clinical prediction guide. In future work, we propose evaluating clinical query filter and machine learning filter models on an independently collected and labeled validation set to generate unbiased results.

Model Generalization

The models generalized with high discriminatory performance in the selected article collections in the purpose, rigor, and format categories. Our results showed that it is possible to build a high performing model if we have a rigorously defined standard with clear semantics. An open theoretical question considers the possibility of building models for article collections that do not have clear selection criteria. Previous work suggested that it is possible to build models with article collections that do not have clear selection criteria. In [14], we built machine learning filter models for an SSOAB gold standard that was constructed without clear selection criteria. Optimizing discriminatory performance for article collections without clear selection criteria is an open area for research. Furthermore, the high discriminatory performance of these models led to a more general observation of methodological rigor. Quality themes and rigor criteria extended across other topics in medicine. Study designs were not specific to one medical topic area. For example, randomized controlled trials were as valid in pediatrics and internal medicine.

Added advantages

The high performance for the machine learning filter models in the clinical prediction guide and economics rigor categories suggest that it is possible to build machine learning filter models for categories that have low positive sample. The positive to negative ratio for economics is 0.069% and the positive to negative ratio for clinical prediction guide is 0.19%. These filters perform well with AUC of 0.997 and 0.985 respectively. Studies of sample size for creating effective models are a useful area of future research.

EBMSearch

The EBMSearch system is a proof of concept system implementing the machine learning filter models. In this paper, we do not explore issues regarding choosing a score above which to show the results or the best presentation for the results. In future work, we would explore methods for choosing scores for the resulting articles that optimize the sensitivity and specificity tradeoff. Also, we would explore the optimal presentation of the results. Web search engines have shown ranked listings to be an effective means to meet an information need. It would be interesting to explore at what point showing articles in dated order is preferable to showing articles in ranked order.

Limitations

The performance of the clinical query filters may be biased. Ideally, within the cross validation design, we would keep the training and testing sets separate. The data used to build the Boolean query should not be the same as that used to evaluate the Boolean query. In this study, the training data used to build the Boolean queries has overlap with the testing data used to test the Boolean queries. Thus clinical query filters may perform better in this study than in an independent test set. In contrast, the design for evaluating performance of the machine learning filter models is a nested cross-validation design which keeps the training data separate from the testing data and avoids bias in performance estimation.

For this study, we chose this performance estimation design for the clinical query filters and machine learning filter models to compare the state of the art available in ranking/ identifying high quality literature.

A potential limitation of any information retrieval study is the choice of gold standard. We selected this gold standard because of the well-documented and rigorous methodology used by Haynes and colleagues to build this standard.

Finally, we do not address the true utility of this system to answer clinical questions effectively or influence medical decision making and outcomes. Evaluating this system and establishing the question types that the system can answer and the impact of the returned results on medical decision making and outcomes are an area for future research.

Conclusions

We have built models that have high discriminatory performance in identifying articles selected by rigorously defined criteria in purpose, format, and rigor categories. These models encompass areas outside of internal medicine and include cost, economics, and clinical prediction guide categories. These models have better or similar discriminatory performance when compared to available clinical query filters. We have also presented a working proof of concept system for implementing these models. This work paves the way for practical application of machine learning filter models to identify high quality articles in the literature.

References

[1] Coiera, E., *Information Economics and the Internet.* J Amer Med Inform Assoc, 2000. 7(3): p. 215-21.

[2] Aphinyanaphongs, Y., et al., *Text Categorization Models for High Quality Article Retrieval in Internal Medicine.* J Amer Med Inform Assoc., 2005. 12(2): p. 207-216.

[3] Wilczynski, N.L. and R.B. Haynes, *Robustness of empirical search strategies for clinical content in MEDLINE.* Proc AMIA Symp, 2002: p. 904-8.

[4] ACP_Journal, *Purpose and Procedure.* ACP Journal, 1999. 131(1): p. A-15 - A-16.

[5] Clinical_Queries. [cited; Available from: http://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.html.

[6] Wilczynski, N. and B. Haynes. *Developing Optimal Search Strategies for Detecting Clinically Sound Causation Studies in MEDLINE*. in *Proceedings AMIA Symposium*. 2003. Washington DC.

[7] Wong, S., et al. *Developing Optimal Search Strategies for Detecting Sound Clinical Prediction Studies in MEDLINE*. in *Proceedings of AMIA Symposium*. 2003. Washington DC.

[8] Robinson, K.A. and K. Dickersin, *Development of highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed.* International Epidemiological Association, 2002. 31: p. 150-153.

[9] Nwosu, C., K. Khan, and P. Chien, *A Two-Term MEDLINE Search Strategy for Identifying Randomized Trials in Obstetrics and Gynecology.* Obstetrics and Gynecology, 1998. 91(4).

[10] Shojania, K.G. and L.A. Bero, *Taking Advantage of the Explosion of Systematic Reviews: An Efficient MEDLINE Search Strategy.* Effective Clinical Practice, 2001. 4(4): p. 157-159.

[11] Bachmann, L.M., et al., *Identifying Diagnostic Studies in MEDLINE: Reducing the Number Needed to Read.* JAMIA, 2002. 9(6): p. 653-658.

[12] *SSOAB.* 12-05-2005 [cited 2005 2005]; Available from: http://www.surgonc.org.

[13] Bernstam, E.V., et al., *Using citation data to improve retrieval from MEDLINE.* J Am Med Inform Assoc, 2006. 13(1): p. 96-105.

[14] Aphinyanaphongs, Y., A. Statnikov, and C.F. Aliferis, *A Comparison of Citation Metrics to Machine Learning Filters for the Identification of High Quality MEDLINE Documents.* J Am Med Inform Assoc, 2006.

[15] Perkovic, D. *Keeping up with the recent research.* 2006 [cited 2007 04-03]; Available from: http://googleblog.blogspot.com/2006/04/keeping-up-with-recent-research.html.

[16] Garfield, E., ed. *Can citation indexing be automated?* Statistical Assocation methods for mechanized documentation, Washington, CD, 1964, ed. M. Stevens, V. Guiliano, and L. Heilprin. 1965: Washington, DC: National Bureau of Standards. 189-192.

[17] Aphinyanaphongs, Y. and C.F. Aliferis, *Text Categorization Models for Identifying Unproven Cancer Treatments on the Web*, in *Medinfo 2007*. 2007: Sydney, Australia.

[18] Fricke, M., et al., *Consumer health information on the Internet about carpal tunnel syndrome: indicators of accuracy.* Am J Med, 2005. 118(2): p. 168-74.

[19] Griffiths, K.M., et al., *Automated assessment of the quality of depression websites.* J Med Internet Res, 2005. 7(5): p. e59.

[20] Tang, T.T., et al., *Quality and Relevance of Domain-specific Search: A Case Study in Mental Health.* Information Retrieval, 2006. 9(2): p. 207-225.

[21] Aphinyanaphongs, Y. and C.F. Aliferis. *Prospective Validation of Text Categorization Filters for Identifying High-Quality Content-Specific Articles in MEDLINE*. in *Proceedings AMIA Symposium*. 2006. Washington DC.

[22] Aphinyanaphongs, Y., R.N. Jerome, and C.F. Aliferis, *Formative Comparative Evaluation of Traditional and Recent Quality-Content Filters for Answering Clinical Questions with MEDLINE*, in *MLA 2007*. 2007: Philadelphia, PA.

[23] Salton, G. and C. Buckley, *Term weighting approaches in automatic retrieval.* Information Processing and Management, 1988. 24(5): p. 513-523.

[24] Porter, M.F., *An algorithm for suffix stripping*. Program, 1980. 14(3): p. 130-137.

[25] *MEDLINE Stopwords*. 3-13-2006 [cited; Available from: http://biolib.princeton.edu/instruct/MedSW.html.

[26] Leopold, E. and J. Kindermann, *Text Categorization with Support Vector Machines. How to Represent Texts In Input Space?* Machine Learning, 2002. 46: p. 423-444.

[27] Vapnik, V., *Statistical Learning Theory*. 1998, New York: Wiley.

[28] Joachims, T. *Text Categorization With Support Vector Machines: Learning With Many Relevant Features*. in *Proceedings of the 10th European Conference On Machine Learning*. 1998: Springer-Velag.

[29] Chang, C. and L. C. *LIBSVM: a library for support vector machines.* 3-13-2006 [cited; Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[30] Morik, K., P. Brockhausen, and T. Joachims. *Combining statistical learning with a knowledge-based approach - A case study in intensive care monitoring*. in *Proc. 16th Int'l Conf. on Machine Learning (ICML-99)*. 1999.

[31] Clinical_Queries. 03-13-2006 [cited; Available from: http://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.html.

[32] Wilczynski, N. and B. Haynes, *Optimal Search Strategies for Detecting Clinically Sound Prognostic Studies in EMBASE.* J Amer Med Inform Assoc., Jul/Aug 2005. 12(4): p. 481-485.

[33] Haynes, B. and N. Wilczynski, *Optimal Search Strategies for retrieving scientifically strong studies of diagnosis from MEDLINE: an analytical survery.* BMJ, 2004.

[34] Weiss, S. and C.A. Kulikowski, *Computer Systems that Learn*. 1991, San Mateo, CA: USA Morgan Kauffman.

[35] Scheffer, T., *Error estimation and model selection*, in *School of Computer Science*. 1999: Technischen Universit at Berlin.

[36] Dudoit, S. and M.J. Van Der Laan, *Asymptotics of cross-validated risk estimation in model selection and performance assessment.* 2003, U.C. Berkeley Division of Biostatistics.

[37] Statnikov, A., et al., *GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data.* International Journal of Medical Informatics, 2005(74): p. 491-503.

[38] *Search Field Descriptions and Tags*. 2007 [cited 2007 04-03]; Available from: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.section.pubmedhelp.Search_Field_Descrip.

[39] *Pubmed's Automatic Term Mapping Enhanced.* NLM Technical Bulletin, 2004(341): p. e7.

[40] Anderson, C., *The Long Tail*. 2006: Hyperion.

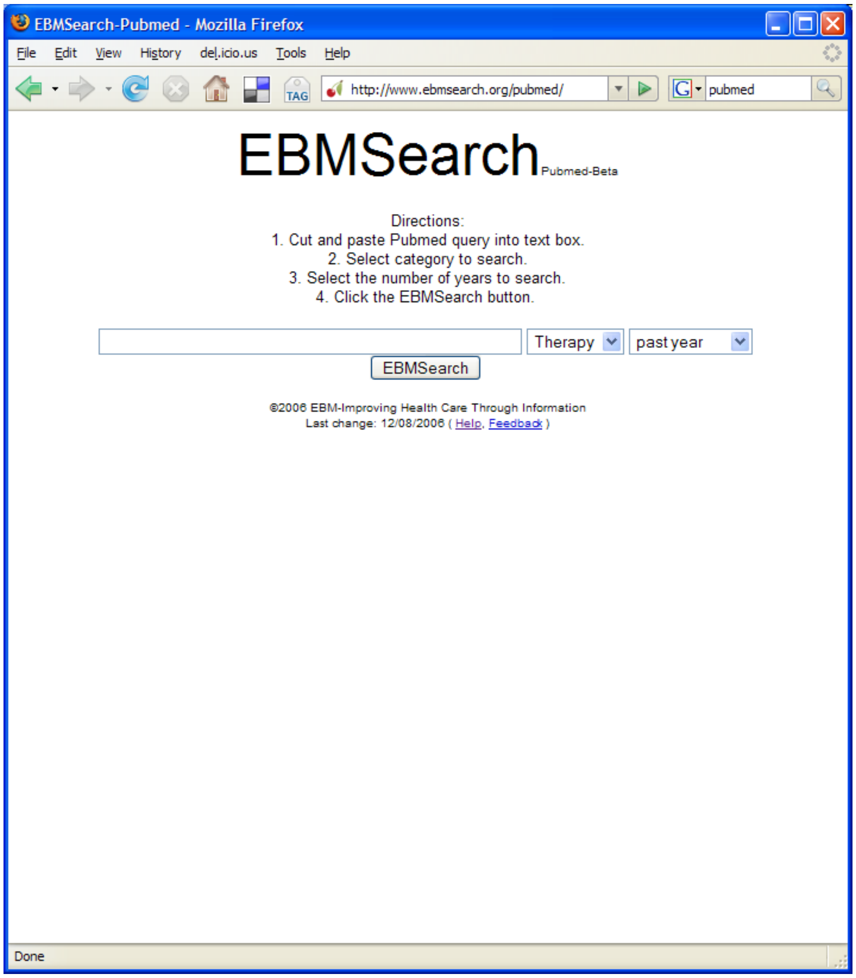
CHAPTER V

V. EBMSEARCH: PROOF OF CONCEPT SEARCH ENGINE

In this section, I present the EBMSearch system that implements the machine learning filter models to all of MEDLINE. In this section, I describe EBMSearch and its implementation.

EBMSearch

I implemented a demonstration proof of concept system called EBMSearch located at http://www.ebmsearch.org. EBMSearch applies machine learning filter models built using the ACP Journal Club gold standard to articles in MEDLINE published from 2000 to 2006*. Users select machine learning filter models built using the ACP Journal Club gold standard in the categories of diagnosis, prognosis, etiology, and treatment, and time frames of 1 year, 2 years, and 5 years for ranking the returned results. The resulting list of articles is ranked by SVM model output score. The index page is shown in Figure V-1.



EBMSearch Pubmed-Beta

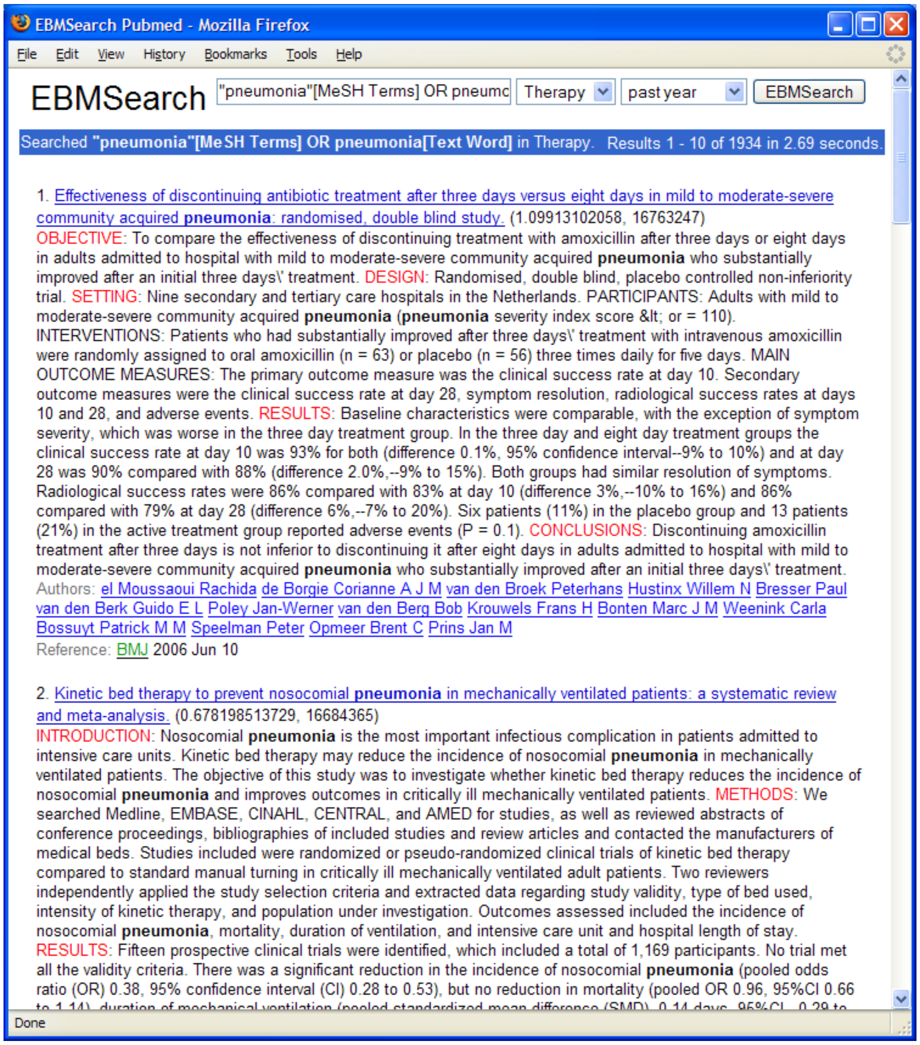
Directions:
1. Cut and paste Pubmed query into text box.
2. Select category to search.
3. Select the number of years to search.
4. Click the EBMSearch button.

©2006 EBM-Improving Health Care Through Information
Last change: 12/08/2006 (Help, Feedback)

Figure V-1 – Index page for EBMSearch system.

The EBMSearch query mirrors the functionality of a Pubmed query. EBMSearch supports field descriptions and tagged search by text word and location (i.e. title, abstract, etc), author, journal title, date, phrase, gender, language, age group, or human or animal studies [1]. EBMSearch also supports automatic term mapping for untagged terms to match, in order, terms in a MeSH translation table, a Journals translation table, Full Author translation table, and an Author index [2]. Any Pubmed query is an acceptable EBMSearch query. In addition, the user specifies a search category (i.e. treatment, diagnosis, prognosis, etiology) and a time frame to search (i.e. 1 year, 2 years, 5 years).

EBMSearch sorts the returned articles by score from the category specific SVM model. Scores above 0 are in the positive class. Scores at 0 are indeterminate, and scores below 0 are in the negative class. The score threshold set at 0 for article classification (i.e. positives/ negatives) is determined by the distribution of positives and negatives in the gold standard for classifying articles across all topics. The classification threshold can be varied based on the user defined threshold of sensitivity and specificity he or she is willing to review. For example, a user writing a review article may read more articles and accept a lower threshold to identify articles with higher sensitivity and lower specificity. A second user may want to read fewer articles and accept a higher threshold to identify articles with lower sensitivity and higher specificity. All articles from the query are sorted and shown from highest to lowest score regardless of relative value. Ranking has proven popular with other major search engines including Pubmed, Yahoo, Google, and MSNSearch. The results page is shown in Figure V-2.



Searched **"pneumonia"[MeSH Terms] OR pneumonia[Text Word]** in Therapy. Results 1 - 10 of 1934 in 2.69 seconds.

1. Effectiveness of discontinuing antibiotic treatment after three days versus eight days in mild to moderate-severe community acquired **pneumonia**: randomised, double blind study. (1.09913102058, 16763247)
OBJECTIVE: To compare the effectiveness of discontinuing treatment with amoxicillin after three days or eight days in adults admitted to hospital with mild to moderate-severe community acquired **pneumonia** who substantially improved after an initial three days\' treatment. DESIGN: Randomised, double blind, placebo controlled non-inferiority trial. SETTING: Nine secondary and tertiary care hospitals in the Netherlands. PARTICIPANTS: Adults with mild to moderate-severe community acquired **pneumonia** (**pneumonia** severity index score &lt; or = 110). INTERVENTIONS: Patients who had substantially improved after three days\' treatment with intravenous amoxicillin were randomly assigned to oral amoxicillin (n = 63) or placebo (n = 56) three times daily for five days. MAIN OUTCOME MEASURES: The primary outcome measure was the clinical success rate at day 10. Secondary outcome measures were the clinical success rate at day 28, symptom resolution, radiological success rates at days 10 and 28, and adverse events. RESULTS: Baseline characteristics were comparable, with the exception of symptom severity, which was worse in the three day treatment group. In the three day and eight day treatment groups the clinical success rate at day 10 was 93% for both (difference 0.1%, 95% confidence interval--9% to 10%) and at day 28 was 90% compared with 88% (difference 2.0%,--9% to 15%). Both groups had similar resolution of symptoms. Radiological success rates were 86% compared with 83% at day 10 (difference 3%,--10% to 16%) and 86% compared with 79% at day 28 (difference 6%,--7% to 20%). Six patients (11%) in the placebo group and 13 patients (21%) in the active treatment group reported adverse events (P = 0.1). CONCLUSIONS: Discontinuing amoxicillin treatment after three days is not inferior to discontinuing it after eight days in adults admitted to hospital with mild to moderate-severe community acquired **pneumonia** who substantially improved after an initial three days\' treatment.
Authors: el Moussaoui Rachida de Borgie Corianne A J M van den Broek Peterhans Hustinx Willem N Bresser Paul van den Berk Guido E L Poley Jan-Werner van den Berg Bob Krouwels Frans H Bonten Marc J M Weenink Carla Bossuyt Patrick M M Speelman Peter Opmeer Brent C Prins Jan M
Reference: BMJ 2006 Jun 10

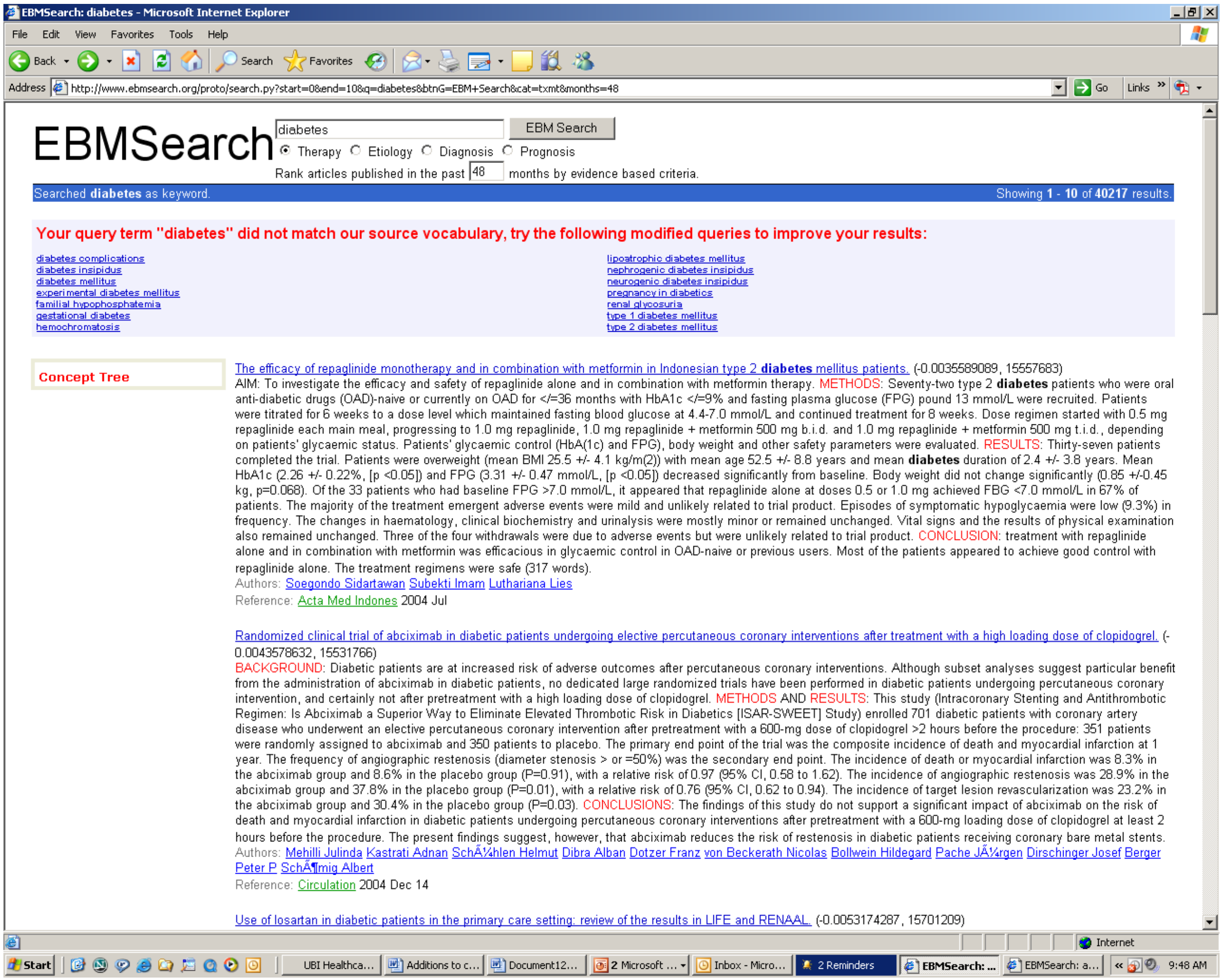
2. Kinetic bed therapy to prevent nosocomial **pneumonia** in mechanically ventilated patients: a systematic review and meta-analysis. (0.678198513729, 16684365)
INTRODUCTION: Nosocomial **pneumonia** is the most important infectious complication in patients admitted to intensive care units. Kinetic bed therapy may reduce the incidence of nosocomial **pneumonia** in mechanically ventilated patients. The objective of this study was to investigate whether kinetic bed therapy reduces the incidence of nosocomial **pneumonia** and improves outcomes in critically ill mechanically ventilated patients. METHODS: We searched Medline, EMBASE, CINAHL, CENTRAL, and AMED for studies, as well as reviewed abstracts of conference proceedings, bibliographies of included studies and review articles and contacted the manufacturers of medical beds. Studies included were randomized or pseudo-randomized clinical trials of kinetic bed therapy compared to standard manual turning in critically ill mechanically ventilated adult patients. Two reviewers independently applied the study selection criteria and extracted data regarding study validity, type of bed used, intensity of kinetic therapy, and population under investigation. Outcomes assessed included the incidence of nosocomial **pneumonia**, mortality, duration of ventilation, and intensive care unit and hospital length of stay. RESULTS: Fifteen prospective clinical trials were identified, which included a total of 1,169 participants. No trial met all the validity criteria. There was a significant reduction in the incidence of nosocomial **pneumonia** (pooled odds ratio (OR) 0.38, 95% confidence interval (CI) 0.28 to 0.53), but no reduction in mortality (pooled OR 0.96, 95%CI 0.66 to 1.14), duration of mechanical ventilation (pooled standardized mean difference (SMD) -0.14 days, 95%CI -0.29 to

Figure V-2 – Results page of EBMSearch system.

I built a results interface that allows quick content skimming of the result set. Query terms are presented in bold in the abstract and/or the title. Major sections in the abstract such as the Introduction, Methods, Results, Conclusions, etc are highlighted in red. Originating journal is highlighted in green. I also used color and highlighting to differentiate parts of the results page. A blue bar separates the search interface from the result articles. Determining whether these search refinements improve skimming of the results page was not pursued in this dissertation.

I implemented EBMSearch using a python based web framework called Turbogears [3]. Turbogears adheres to a Model-View-Controller architecture which has advantages in separating data, logic, and presentation of the website [4]. Whether this system would function with production like loads is unknown.

In earlier versions, I implemented a spell checker, an automated suggestion list based on MeSH terms to help users refine their queries, and semi-automatic MeSH explosions. These GUI features were excluded from the current version. The spell checker would identify misspellings using the Aspell algorithm [5]. The automated suggestion list proposed MeSH terms that included the query terms. We implemented the suggestion list by adding terms from MeSH and synonyms to an inverted index [6]. For example, given the query “diabetes” the system would suggest MeSH term “diabetes mellitus type II” since the term diabetes occurs in the MeSH term. Similarly, for a query such as “heart attack”, the system would suggest “myocardial infarction” as the appropriate MeSH descriptor to refine the results. We obtained the MeSH terms and synonyms from download files available from the National Library of Medicine [7]. I also considered semi-automatic MeSH explosions. For example, a search for “pneumonia” would automatically map to “Pneumonia [MAJR]”. Pubmed has since implemented a spell checker [8] and MeSH explosions for query refinement [9]. Research supporting these user interface changes is a logical step for future research. An earlier version of the results page showing some of these features is shown in Figure V-3.



http://www.ebmsearch.org/proto/search.py?start=0&end=10&q=diabetes&btnG=EBM+Search&cat=txmt&months=48

EBMSearch

diabetes [EBM Search]

● Therapy ○ Etiology ○ Diagnosis ○ Prognosis

Rank articles published in the past 48 months by evidence based criteria.

Searched **diabetes** as keyword. Showing **1** - **10** of **40217** results.

Your query term "diabetes" did not match our source vocabulary, try the following modified queries to improve your results:

diabetes complications
diabetes insipidus
diabetes mellitus
experimental diabetes mellitus
familial hypophosphatemia
gestational diabetes
hemochromatosis
lipoatrophic diabetes mellitus
nephrogenic diabetes insipidus
neurogenic diabetes insipidus
pregnancy in diabetics
renal glycosuria
type 1 diabetes mellitus
type 2 diabetes mellitus

Concept Tree

The efficacy of repaglinide monotherapy and in combination with metformin in Indonesian type 2 **diabetes** mellitus patients. (-0.0035589089, 15557683)
AIM: To investigate the efficacy and safety of repaglinide alone and in combination with metformin therapy. METHODS: Seventy-two type 2 **diabetes** patients who were oral anti-diabetic drugs (OAD)-naive or currently on OAD for </=36 months with HbA1c </=9% and fasting plasma glucose (FPG) pound 13 mmol/L were recruited. Patients were titrated for 6 weeks to a dose level which maintained fasting blood glucose at 4.4-7.0 mmol/L and continued treatment for 8 weeks. Dose regimen started with 0.5 mg repaglinide each main meal, progressing to 1.0 mg repaglinide, 1.0 mg repaglinide + metformin 500 mg b.i.d. and 1.0 mg repaglinide + metformin 500 mg t.i.d., depending on patients' glycaemic status. Patients' glycaemic control (HbA(1c) and FPG), body weight and other safety parameters were evaluated. RESULTS: Thirty-seven patients completed the trial. Patients were overweight (mean BMI 25.5 +/- 4.1 kg/m(2)) with mean age 52.5 +/- 8.8 years and mean **diabetes** duration of 2.4 +/- 3.8 years. Mean HbA1c (2.26 +/- 0.22%, [p <0.05]) and FPG (3.31 +/- 0.47 mmol/L, [p <0.05]) decreased significantly from baseline. Body weight did not change significantly (0.85 +/-0.45 kg, p=0.068). Of the 33 patients who had baseline FPG >7.0 mmol/L, it appeared that repaglinide alone at doses 0.5 or 1.0 mg achieved FBG <7.0 mmol/L in 67% of patients. The majority of the treatment emergent adverse events were mild and unlikely related to trial product. Episodes of symptomatic hypoglycaemia were low (9.3%) in frequency. The changes in haematology, clinical biochemistry and urinalysis were mostly minor or remained unchanged. Vital signs and the results of physical examination also remained unchanged. Three of the four withdrawals were due to adverse events but were unlikely related to trial product. CONCLUSION: treatment with repaglinide alone and in combination with metformin was efficacious in glycaemic control in OAD-naive or previous users. Most of the patients appeared to achieve good control with repaglinide alone. The treatment regimens were safe (317 words).
Authors: Soegondo Sidartawan Subekti Imam Luthariana Lies
Reference: Acta Med Indones 2004 Jul

Randomized clinical trial of abciximab in diabetic patients undergoing elective percutaneous coronary interventions after treatment with a high loading dose of clopidogrel. (-0.0043578632, 15531766)
BACKGROUND: Diabetic patients are at increased risk of adverse outcomes after percutaneous coronary interventions. Although subset analyses suggest particular benefit from the administration of abciximab in diabetic patients, no dedicated large randomized trials have been performed in diabetic patients undergoing percutaneous coronary intervention, and certainly not after pretreatment with a high loading dose of clopidogrel. METHODS AND RESULTS: This study (Intracoronary Stenting and Antithrombotic Regimen: Is Abciximab a Superior Way to Eliminate Elevated Thrombotic Risk in Diabetics [ISAR-SWEET] Study) enrolled 701 diabetic patients with coronary artery disease who underwent an elective percutaneous coronary intervention after pretreatment with a 600-mg dose of clopidogrel >2 hours before the procedure: 351 patients were randomly assigned to abciximab and 350 patients to placebo. The primary end point of the trial was the composite incidence of death and myocardial infarction at 1 year. The frequency of angiographic restenosis (diameter stenosis > or =50%) was the secondary end point. The incidence of death or myocardial infarction was 8.3% in the abciximab group and 8.6% in the placebo group (P=0.91), with a relative risk of 0.97 (95% CI, 0.58 to 1.62). The incidence of angiographic restenosis was 28.9% in the abciximab group and 37.8% in the placebo group (P=0.01), with a relative risk of 0.76 (95% CI, 0.62 to 0.94). The incidence of target lesion revascularization was 23.2% in the abciximab group and 30.4% in the placebo group (P=0.03). CONCLUSIONS: The findings of this study do not support a significant impact of abciximab on the risk of death and myocardial infarction in diabetic patients undergoing percutaneous coronary interventions after pretreatment with a 600-mg loading dose of clopidogrel at least 2 hours before the procedure. The present findings suggest, however, that abciximab reduces the risk of restenosis in diabetic patients receiving coronary bare metal stents.
Authors: Mehilli Julinda Kastrati Adnan SchÃ¼hlen Helmut Dibra Alban Dotzer Franz von Beckerath Nicolas Bollwein Hildegard Pache JÃ¼rgen Dirschinger Josef Berger Peter P SchÃ¶mig Albert
Reference: Circulation 2004 Dec 14

Use of losartan in diabetic patients in the primary care setting: review of the results in LIFE and RENAAL. (-0.0053174287, 15701209)

Figure V-3 – Proof of concept interface. The left side will display the MeSH tree if one of the terms is a MeSH term. Suggestions are made if the search query matches part of the MeSH term vocabulary. For example, “diabetes mellitus” and “diabetes insipidus” are suggested for a search of “diabetes.”

I excluded the earlier version with these GUI enhancements from production. Their implementation is not within the present focus of this proof of concept system[8]. In future work, I will create a more robust search engine for general use.

References

[1] Search Field Descriptions and Tags. 2007 [cited 2007 04-03]; Available from: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.section.pubmedhelp.Search_Field_Descrip

[2] Pubmed's Automatic Term Mapping Enhanced. NLM Technical Bulletin. 2004 Nov-Dec(341):e7.

[3] Turbogears. [cited 2007 7-5]; Available from: http://www.turbogears.org

[4] Deacon J. Model - View - Controller (MVC) Architecture. [cited 2007 7-5]; Available from: http://www.jdl.co.uk/briefings/MVC.pdf

[5] Aspell. [cited 2007 7-4]; Available from: http://aspell.net

[6] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. Harlow, England: Addison-Wesley 1999.

[7] Medical Subject Headings. [cited 2007 04-03]; Available from: http://www.nlm.nih.gov/mesh/filelist.html

[8] Canese K. New Spell Checking Feature. NLM Technical Bulletin. 2004 Nov-Dec;341:e12.

[9] Knecht L, Canese K, Port T. Pubmed: Truncation, Automatic Explosion, Mapping, and MeSH Headings. NLM Technical Bulletin. 1998 May - June;302.

[8] Pubmed has since introduced a spell checker and semi-automatic MeSH explosions.

CHAPTER VI

VI. EXTENSIONS TO THE WORLD WIDE WEB

In this section, I extend the machine learning framework to the web. I built the first validated models that identify web pages that make unproven cancer treatment claims outperforming unvalidated web models and PageRank by 30% area under the receiver operating curve.

Text Categorization Models for Identifying Unproven Cancer Treatments on the Web

Aphinyanaphongs, Y, Aliferis C. “Text Categorization Models for Identifying Unproven Cancer Treatments on the Web.” In: Medinfo 2007; Sydney, Australia.

Abstract

The nature of the internet as a non-peer-reviewed (and largely unregulated) publication medium has allowed wide-spread promotion of inaccurate and unproven medical claims in unprecedented scale. Patients with conditions that are not currently fully treatable are particularly susceptible to unproven and dangerous promises about miracle treatments. In extreme cases, fatal adverse outcomes have been documented. Most commonly, the cost is financial, psychological, and delayed application of imperfect but proven scientific modalities. To help protect patients, who may be

desperately ill and thus prone to exploitation, we explored the use of machine learning techniques to identify web pages that make unproven claims. This feasibility study shows that the resulting models can identify web pages that make unproven claims in a fully automatic manner, and substantially better than previous web tools and state-of-the-art search engine technology.

Introduction

"The killing of all parasites and their larval stages together with removal of isopropyl alcohol and carcinogens from the patients' lifestyle results in remarkable recovery (from cancer), generally noticeable in less than one week [1]." This is one example of an unproven treatment claim made on the web. These unproven treatments are known as *quackery* with the *quacks* promoting them defined as "untrained people who pretend to be physicians and dispense medical advice and treatment [2]." The internet allows quacks to advocate inaccurate and unproven treatments with documented fatal, adverse outcomes in some situations [3-6].

In regards to cancer patients, Metz et al. reported that 65% of cancer patients searched unproven treatments and 12% purchased unconventional medical therapies online [7]. In another study, Richardson reported that 83% of cancer patients had used at least one unproven treatment [8]. Many patients are ill-equipped to evaluate treatment information [9]. The language and quality of web pages with unproven treatments is also highly variable [10]. The rapid growth of the internet, combined with the ease of publishing unproven claims leads susceptible and often desperately ill patients to further adverse outcomes, patient and family despair, and sunk costs. It is thus an important mandate of the medical profession to protect patients from inaccurate and poor medical information.

So far extensive research has developed several manual methods to combat the propagation of unproven claims on the web. The Health-on-the-Net Foundation advocates self-regulation of health related websites [11]. The foundation applies strict criteria to websites and grants them a seal of approval if they pass. However, most health care consumers ignore the seals [12]. In another approach, experts produced rating tools that consumers are supposed to apply to websites[13, 14]. Another method is manual review of individual websites that are published either in print or electronically.

Each method has limitations. Self-regulation relies on knowledge of the certification and a vigilant public to report failing web sites. Rating tools are dependent on a knowledgeable public to apply, they are difficult to validate, time consuming to produce, and do not always produce consistent ratings [15, 16]. Moreover, the rating tools are not appropriate for use on complementary/ alternative medicine sites [17]. Furthermore, manual review suffers from limits in reviewer time and the selection of web sites to review.

Ideally, we would like a solution that is validated, easy to apply by health care consumers, and works on any webpage. In this paper, we hypothesize that automated approaches to identifying web pages with unproven claims may provide a solution.

Previous Work On Automatic Webpage Identification

Previous research focused on automated or semi-automated approaches to identifying *high quality* medical web pages.

Price and Hersh [18] evaluated web page content by combining a score measuring quality proxies for each page. Quality proxies included relevance, credibility, bias, content, currency, and the value of its links. The authors evaluated the algorithm on a small test collection of 48 web pages covering nine medical topics labeled as desirable or undesirable by the investigator. In all cases, the score assigned to the desirable pages was higher than the scores assigned to undesirable pages.

Even though the algorithm perfectly discriminated between desirable and undesirable webpages, several limitations exist. The test sample was small and not representative of the scale for a web classification task. The algorithm does not measure content quality directly, but used proxies for quality to compile a score for a web page. The usefulness of some of the explicit criteria may not correlate with content quality [19], and may not be valid or good features to include for scoring.

As a leading search engine, Google has become a de facto standard for identifying and ranking web pages. Pages that rank highly in Google are assumed to be of better quality than those at lower rank. Several researchers have explored this assumption for health pages. Fricke and Fallis [20] evaluated PageRank score as one indicator of quality for 116 web sites about carpal tunnel syndrome. Their results show that PageRank score is not inherently useful for discrimination or helping users to avoid inaccurate or poor information. Of the 70 web sites with high PageRank, 29 of them had inaccurate information.

Griffiths [21] evaluated PageRank scores for depression websites using evidence based quality scores. The authors obtained Google PageRank scores for 24 depression websites from the DMOZ Open Directory Project website. Two health professional raters assigned an evidence based quality score to each site. PageRank scores correlated weakly (r = 0.61, P=0.002) with the evidence based quality scores.

Tang, Craswell, and Hawking [22] compared Google results with a domain-specific search engine for depression. They found that of a 101 selected queries, Google returned more relevant results, but at the expense of quality. Of the 50 treatment related queries, Google returned 70 pages of which 19 strongly disagreed with the scientific evidence.

Hypothesis

Our fundamental hypothesis for this feasibility study is that we can model expert opinion and build machine learning models that identify web pages that make unproven claims for the treatment of cancer.

To the best of our knowledge, there is no research on validated automated techniques for identifying web pages that make unproven claims. In prior work, we showed that text categorization methods identified high quality content specific articles in internal medicine [23]. Extending this work into the web space, we reverse the hypothesis of the previous studies. Rather than identifying high quality pages, we explore automated identification of low quality pages, specifically pages that make unproven claims for cancer treatment.

Materials and Methods

Definitions

Our gold standard relied on selected unproven cancer treatments identified by experts at http://www.quackwatch.org. The website is maintained by a 36 year old nonprofit organization whose mission is to "combat health related frauds, myths, fads, fallacies, and misconduct." The group employs a 152 person scientific and technical advisory board composed of academic and private physicians, dentists, mental health advisors, registered dietitians, podiatrists, veterinarians, and other experts whom review health related claims. By using unproven treatments identified by an oversight organization, we capitalized on an existing high quality review.

Corpus Construction

For this feasibility study, we randomly chose 8 unproven treatments from 120 dubious cancer treatments listed by quackwatch.org [24]. The randomly selected treatments were "Cure for all Cancers", "Mistletoe", "Krebiozen", "Metabolic Therapy", "Cellular Health", "ICTH", "Macrobiotic Diet", and "Insulin Potentiation Therapy." We then identified web pages that have these treatments by appending the words "cancer" and "treatment" and querying Google. We retrieved the top 30 results for each unproven treatment. We used a python script to download and store each result as raw html for further labeling.

Corpus Labels

We applied a set of criteria for identifying web pages with unproven treatment claims. First, of the initial 240 pages, we excluded not found (404 response code) error pages, no content pages, non-English pages, password-protected pages, pdf pages, redirect pages, and pages where the actual treatment text does not appear in the document[9]. Of the remaining 191 html pages, both authors independently asked the following question of each web page: does the web page make unproven claims about the proposed treatment and its efficacy. We labeled web pages with unproven claims as positive and the others as negative.

Web pages that are purely informational in nature but do not make any unproven claims about the cancer treatment and its efficacy were labeled as negative. Web pages selling a book with user comments that has unproven claims were labeled as positive. Portal pages that do not make any claim were labeled as negative. Web pages that attempted to present an objective viewpoint of the treatment were carefully reviewed for any unproven claims, and, if so, were labeled positive. Additionally web pages that sell unproven treatments but do not make claims were labeled negative.

Both authors applied the criteria independently. We calculated the inter-observer agreement (Cohen's Kappa [25]) at 0.76[10]. Of the 20 sites with discrepant labelings, the reviewers discussed the labels until consensus was reached. The final corpus was composed of 191 web pages with 93 labeled as positive and 98 as negative.

[9] The Google ranking algorithm relies on anchor text to identify web page content. Anchor text may point to a web page that does not use the anchor text in the web page itself.

[10] We set a threshold of 0.70 for Cohen's Kappa. If kappa was below 0.70, we would refine the labeling criteria until the threshold was reached.

Webpage Preparation

For this feasibility study, we chose the simplest web page representation. We converted web pages to a "bag of words" suitable for the machine learning algorithm[23]. First, for each web page, we removed all content between style and script tags. Second, all tags (including the style and script tags) were removed. Third, we replaced all punctuation with spaces. We split the remaining string on the spaces to obtain individual words. Finally, we stemmed each word [23], applied a stop word list [23], removed any words that appear in less than 3 web pages, and encoded as weighted features using a log frequency with redundancy scheme [23].

Learning Model (Support Vector Machines)

We employed Support Vector Machine (SVM) classification algorithms. The SVM's calculate maximal margin hyperplane(s) separating two or more classes of the data. SVMs have had superior text classification performance compared to other methods [23], and this motivated our use of them. We used an SVM classifier implemented in libSVM v2.8 [26] with a polynomial kernel. We optimized the SVM penalty parameter C over the range {0.1, 1, 2, 5, 10} with imbalanced costs applied to each class proportional to the priors in the data [23], and degree d of the polynomial kernel over the range {1, 2, 5}. The ranges of costs and degrees for optimization were chosen based on previous empirical studies [23].

Model Selection and Performance Estimation

We used 10-fold cross-validation that provides unbiased performance estimates of the learning algorithms [23]. This choice for n provided sufficient high-quality positive samples for training in each category and provided sufficient article samples for the classifiers to learn the models. The cross-validation procedure first divided the data randomly into 10 non-overlapping subsets of documents where the proportion of positive and negative documents in the full dataset is preserved for each subset. Next, the following was repeated 10 times: we used one subset of documents for testing (the "original testing set") and the remaining nine subsets for training (the "original training set") of the classifier. The average performance over 10 original testing sets is reported.

To optimize parameters of the SVM algorithms, we used another "nested" loop of cross-validation [23] by splitting each of the 10 original training sets into smaller training sets and validation sets. For each combination of learner parameters, we obtained cross-validation performance and selected the best performing parameters inside this inner loop of cross-validation. We next built a model with the best parameters on the original training set and applied this model to the original testing set. Notice that the final performance estimate obtained by this procedure will be unbiased because each original testing set is used only once to estimate performance of a single model that was built by using training data exclusively.

Quackometer

We compared our algorithm to a heuristic, unvalidated, and unpublished quack detection tool available at http://www.quackometer.net. The exact details of the detection tool are proprietary. In general, the algorithm counts words in web pages that quacks use, and sorts the words into at least 5 dictionaries [27]. It looks for altmed terms such as "homeopathic" and "herbal", pseudoscientific words such as "toxins" and "superfoods", domain specific words such as "energy" and "vibration", skeptical words such as "placebo" and "flawed", and commerce terms such as "products" and "shipping". The algorithm counts the frequency of terms, applies a frequency threshold, and generates a corresponding score from 0 to 10. The tool is available at [28].

We compared our models to the Quackometer by calculating the corresponding area under the curve (AUC) for each 10 fold-split and reporting the mean and standard deviation.

Google PageRank

The PageRank algorithm [29] is used by Google to identify higher quality pages on the web. The basic tenet is that a web page will rank highly if the web page has more and higher quality links pointing to it. For example, if a web page has a link from Yahoo (a highly linked page), it would rank higher than a link from a less linked to web page. In detecting web pages with unproven claims, our assumption is that web pages with poor quality information should get fewer and lower quality links than web pages with better quality.

We used Google as a proxy for PageRank[11]. We made the comparison to our algorithms within each topic rather than within each 10 fold split. We compared within each topic to avoid bias in ranking situations where one topic has uniformly higher Google rank than another topic. We inverted the labels[12] in the 8 randomly selected topics, calculated the AUC, and reported the mean AUC and standard deviation.

Results

Table VI-1 shows the AUC performance between the machine learning filter models, Quackometer, and Google. The machine learning method identified web pages that make unproven claims with an AUC of 0.93 with a standard deviation of 0.05 across the 10 folds. Quackometer does worse with an AUC of 0.67 and a standard deviation of 0.10 across the same 10 folds. Finally Google performs least effectively in discriminating web pages with an AUC of 0.63 and a standard deviation of 0.17 across the 8 selected topics. Figure VI-1 shows the corresponding receiver operating curves for each method.

[11] Google uses a proprietary version of PageRank for ranking.

[12] We test the assumption that PageRank will rank web pages with *proven* claims higher than web pages with *unproven* claims.

Table VI-1 – Area Under Curve for Each Discrimination Method

Model	**Mean Area Under the Curve**
Support Vector Machine	0.93 (std. 0.05)
Quackometer	0.67 (std. 0.10)
Google	0.63 (std. 0.17)[13]

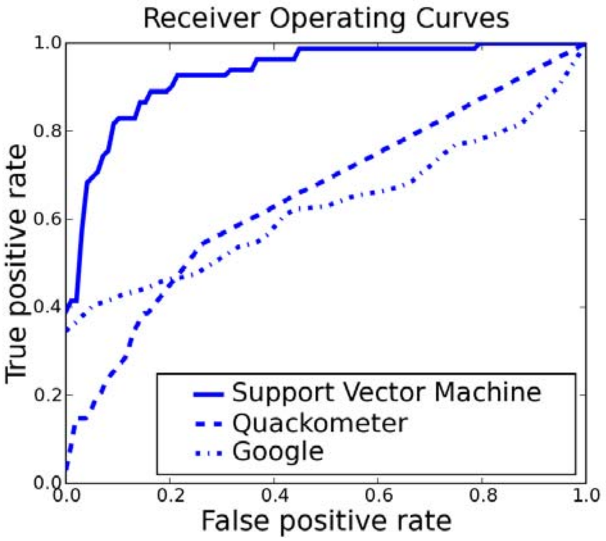


Figure VI-1 - Receiver operating curves for each method.

Discussion

This feasibility study showed that machine learning filter models identify web pages that make unproven claims on a select, focused gold standard. The learning filters have superior performance over the Quackometer [27] and Google. We also note that the loose

[13] The mean and standard deviation are calculated across the 8 topics rather than across the test sets of the 10 folds.

correlation between Google and high quality sites seems comparable to previous work [20-22].

This method has distinct advantages to rating instruments [13, 14] or manual review. First, there is no need to state explicit rating criteria. The model identified patterns in the data that label a page with unproven claims. Second, compared to the limited focus of manual review on select web pages, these models allow application to any web page.

We also highlight a subtle point in this work. We make a distinction between web pages that make unproven claims and web pages that promote the unproven treatment. Oftentimes, this distinction is blurry. For this work, we only want to identify pages that make unproven claims. Pages that promote a product but do not make unproven claims are not identified. In future studies, we are interested in evaluating models that identify web pages that promote treatments.

In Table VI-2, we present excerpts from pages where the previous models failed to identify pages with unproven claims. These pages should have been identified by the Quackometer [27] and should not have appeared in the top 30 Google results. Such failure to identify or mark these pages may result in patient's exposure to potentially harmful, unproven treatments.

In practice, we envision implementing a system that works much like a spam filter works for e-mail. Spam filters identify illegitimate e-mails. In a similar fashion, we envision a system that runs on top of a search engine and flags any web pages that may have unproven health claims.

Table VI-2 - Web page excerpts where previous tools fail to detect unproven claims. A page that makes unproven claims is identified as such if it has a small support vector machine rank, a large quackometer score, and a large Google rank, respectively. SVM rank is calculated over 10 fold cross validation test set composed of 9 positives and 9 negatives. Google rank is out of the top 30 results returned. Quackometer score provides ranks from 0 to 10. "S" denotes success of the corresponding filter, while "F" failure.

Failure Analysis Excerpts	Support Vector Machine Rank	Quackometer score	Google rank
I am convinced that our mind and emotions are the deciding factor in the cure of cancer.	1 (S)	1 (F)	16 (F)
The hundreds of clinical studies conducted by many competent physicians around the world, including those directed by Dr. Emesto Contreras Rodriguez at the Oasis of Hope Hospital hospital in Mexico, give us complete confidence that there is no danger.	3 (S)	0 (F)	9 (F)
The cure shows results almost immediately and lasts three weeks only. It is cheap and affordable for everybody and proved with 138 case studies.	3 (S)	8 (S)	3 (F)
Many advanced cancer patients are petrified of their tumor. This knee-jerk reaction is caused by orthodox medicine's focus on the highly profitable (and generally worthless) process of shrinking tumors.	1 (S)	1 (F)	18 (F)
IPT (Insulin Potentation Therapy) has an outstanding 135 doctor-year track record (115 years for cancer) over 72 years, and is ready for clinical trials and widespread use.	1 (S)	0 (F)	1 (F)
We are proud of these findings, which confirm that cellular medicine offers solutions for the most critical process in cancer development, the invasion of cancer cells to other organs in the body. Conventional medicine is powerless in this.	2 (S)	1 (F)	8 (F)

Limitations

We tested a small sample comprised of 8 unproven treatments in 240 web pages. We will explore how well the models generalize with an independently collected dataset, more unproven treatments, and more labeled web pages. Collecting an independent dataset would allow for validation of the labeling criteria and the model selection procedures. For this feasibility study, we purposely limited the topic of this study to cancer treatment. In the future, we will build and evaluate other models identifying web pages that make unproven claims for other conditions such as arthritis, autism, and allergies.

Conclusions

We present a first of its kind feasibility study to build machine learning filter models that exhibit high discriminatory performance for identifying web pages with unproven cancer treatment claims. This work paves the way for building broadly applicable models involving more health conditions, more pages with unproven claims, and eventually applied systems to protect patients from quackery.

Acknowledgments

The first author acknowledges support from NLM grant LM007948-02. The second author acknowledges support from grant LM007948-01.

References

[1] Clark H. *The Cure for All Cancers*: New Century Press; 1993.

[2] *American Heritage Dictionary*.

[3] Hainer MI, Tsai N, Komura ST, Chiu CL. Fatal hepatorenal failure associated with hydrazine sulfate. *Ann Intern Med*. 2000 Dec 5;133(11):877-80.

[4] See KA, Lavercombe PS, Dillon J, Ginsberg R. Accidental death from acute selenium poisoning. *Med J Aust*. 2006 Oct 2;185(7):388-9.

[5] Bromley J, Hughes BG, Leong DC, Buckley NA. Life-threatening interaction between complementary medicines. *Ann Pharmacother*. 2005 Sep;39(9):1566-9.

[6] Mularski RA, Grazer RE, Santoni L, Strother JS, Bizovi KE. Treatment advice on the internet leads to a life-threatening adverse reaction: hypotension associated with Niacin overdose. *Clin Toxicol (Phila)*. 2006;44(1):81-4.

[7] Metz JM, Devine P, DeNittis A, Jones H, Hampshire M, Goldwein J, Whittington R. A multi-institutional study of Internet utilization by radiation oncology patients. *Int J Radiat Oncol Biol Phys*. 2003 Jul 15;56(4):1201-5.

[8] Richardson MA, Sanders T, Palmer JL, Greisinger A, Singletary SE. Complementary/alternative medicine use in a comprehensive cancer center and the implications for oncology. *J Clin Oncol*. 2000 Jul;18(13):2505-14.

[9] Sagaram S, Walji M, Bernstam E. Evaluating the prevalence, content and readability of complementary and alternative medicine (CAM) web pages on the internet. *Proc AMIA Symp*. 2002:672-6.

[10] Ernst E, Schmidt K. 'Alternative' cancer cures via the Internet? *Br J Cancer*. 2002 Aug 27;87(5):479-80.

[11] Health on the Net.   [accessed 11-27-2006]; http://www.hon.ch/

[12] Eysenbach G, Kohler C. How Do Consumers Search For and Appraise Health Information on the WWW? *BMJ*. 2002 March 9;324.

[13] Bernstam EV, Shelton DM, Walji M, Meric-Bernstam F. Instruments to assess the quality of health information on the World Wide Web. *Int J Med Inform*. 2005 Jan;74(1):13-9.

[14] Kim P, Eng TR, Deering MJ, Maxfield A. Published criteria for evaluating health related web sites: review. *Bmj*. 1999 Mar 6;318(7184):647-9.

[15] Bernstam EV, Sagaram S, Walji M, Johnson CW, Meric-Bernstam F. Usability of quality measures for online health information. *Int J Med Inform*. 2005 Aug;74(7-8):675-83.

[16] Ademiluyi G, Rees CE, Sheard CE. Evaluating the reliability and validity of three tools to assess the quality of health information on the Internet. *Patient Educ Couns*. 2003 Jun;50(2):151-5.

[17] Walji M, Sagaram S, Sagaram D, Meric-Bernstam F, Johnson C, Mirza NQ, Bernstam EV. Efficacy of quality criteria to identify potentially harmful information. *J Med Internet Res*. 2004 Jun 29;6(2):e21.

[18] Price SL, Hersh WR. Filtering Web pages for quality indicators: an empirical approach to finding high quality consumer health information on the World Wide Web. *Proc AMIA Symp*. 1999:911-5.

[19] Fallis D, Fricke M. Indicators of accuracy of consumer health information on the Internet. *J Am Med Inform Assoc*. 2002 Jan-Feb;9(1):73-9.

[20] Fricke M, Fallis D, Jones M, Luszko GM. Consumer health information on the Internet about carpal tunnel syndrome. *Am J Med*. 2005 Feb;118(2):168-74.

[21] Griffiths KM, Tang TT, Hawking D, Christensen H. Automated assessment of the quality of depression websites. *J Med Internet Res*. 2005;7(5):e59.

[22] Tang TT, Craswell N, Hawking D, Griffiths KM, Christensen H. Quality and Relevance of Domain-specific Search: A Case Study in Mental Health. *Info Retr*. 2006;9(2):207-25.

[23] Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text Categorization Models for High Quality Article Retrieval in Internal Medicine. *J Amer Med Inform Assoc*. 2005;12(2):207-16.

[24] Cancer Patients Seeking Alternative Treatments. [accessed 11-26-2006]; http://www.quackwatch.org/00AboutQuackwatch/altseek.html

[25] Cohen J. A coefficient of agreement for nominal scales. *Education and Psych Measurement*. 1960;20(1):37-46.

[26] Chang C, C. L. LIBSVM. 3-13-2006 [accessed; http://www.csie.ntu.edu.tw/~cjlin/libsvm

[27] Science of Quackometrics. [accessed 11-26-2006; http://www.quackometer.net/blog/2006/04/science-of-quackometrics.html

[28] Quackometer. [accessed 11-26-2006]; http://www.quackometer.net/?page=quackometer

[29] Brin S, Page L. The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems*. 1998;30:107-17.

CHAPTER VII

VII. CONCLUSIONS

This dissertation addresses the problem of identifying high quality journal articles and web-sites in support of Evidence-Based Medicine using machine learning methodologies. As the volume and size of the medical literature and the web continue to grow, the need for automated techniques to filter and identify quality literature and web pages becomes paramount.

The experimental evidence amassed and discussed in the previous chapters supports a number of conclusions:

1. The models produced and evaluated have excellent predictivity for identifying high quality articles.
2. The discriminatory ability of machine learning models is superior to Boolean filters, bibliometric citation count, impact factor, Google Pagerank, Yahoo Webranks, and web page hit count.
3. The model selection procedures employed yield models that are not overfit and their performance generalizes well in prospective validation corpora.
4. Models can be built for many medical specialty areas outside of internal medicine.
5. The models can identify quality articles in many content categories.

6. The models are straightforward to implement as demonstrated with a proof of concept web-based system that applies the models to MEDLINE articles.
7. Finally, the methodology was also capable of producing models with high discriminatory performance for identifying web pages that make unproven treatment claims.

The set of hypotheses explored and the experiments presented also point out to several significant open questions. I discuss here several such related problems that are both non-trivial and important to solve.

Open Problems

1. In evaluating the machine learning filter models, I ranked the articles by score and applied receiver operating characteristic curve analysis. The receiver operating characteristic curve shows multiple points of sensitivity and specificity obtained across all topics. Specific users will need the models to rank documents within one or just a few topics, however. Because each topic has prior probabilities of positive to negative articles than the totality of PubMed, topic-specific thresholds will be needed to ensure performance characteristics such as smallest number of documents needed to be read in order to see x% of all positive ones, etc. Such thresholds can be pre-computed for topics and topic categories, or dynamic schemes employing user-feedback may be utilized.
2. Following standard modeling methodology principles, I started with simple representation of articles, that is the "bag of words" (and occasionally the "bag of

concepts") approach before considering more complex representations. The experiments showed that the simple representations were very effective for the tasks studied. In future tasks, especially as the granularity of questions answered by the machine learning models increases, more complex representations may be needed.

3. The machine learning filter models have very high discriminatory power with area under the ROC curves greater than 90% in all categories and close to 99% for many of them. For the categories where area under the curve is in the low 90s (%),in-depth failure analysis on false positives and false negatives may shed light on techniques needed to further improve discriminatory performance.
4. How vulnerable are the machine learning filter models to being gamed? Understanding the mechanisms by which this is feasible and preventing them is an interesting and necessary area of work.
5. The machine learning filter models outperform citation metrics but the former require a labeled gold standard whereas the latter do not. In the majority of experiments presented, pre-existing labeling was used and the manual effort of labeling documents was thus minimized. In information retrieval tasks where existing labels cannot be readily found, it would be valuable to have methods that create such by using/processing citation structure information, by employing imputation or via other semi-supervised methods.
6. An important open question is how the machine learning filter models can help answer clinical questions and influence medical decision making and outcomes.

7. As discussed, I built a proof-of-concept system that demonstrated the application of the machine learning filter models on MEDLINE. The returned results are shown in a list format. An open area of research is to explore state of the art ways to present and highlight the pertinent information from the abstracts.
8. Finally, the experiments to identify unproven cancer treatments in the WWW just scratched the surface of what is possible. Increasing the scope of diseases, treatments, evaluation years, and types of questions asked define a large space of possibilities for health-related information retrieval on the WWW.