Research Paper ■

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

YINDALON APHINYANAPHONGS, MS, MS, ALEXANDER STATNIKOV, MS, MS, CONSTANTIN F. ALIFERIS, MD, PHD

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.

■ **J Am Med Inform Assoc.** 2006;13:446–455. DOI 10.1197/jamia.M2031.

Introduction and 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 reinvigorated 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 potentially 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

Affiliations of the authors: Discovery Systems Laboratory, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN

Mr. Aphinyanaphongs was supported by NLM grant LM007948-02 and the Medical Library Association's Donald Lindberg Research Fellowship. Dr. Aliferis was funded by 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.

Correspondence and reprints: Constantin F. Aliferis, MD, PhD, Department of Biomedical Informatics, Eskind Biomedical Library, room 412, Vanderbilt University, 2209 Garland Avenue, Nashville, TN 37232; e-mail: ⟨constantin.aliferis@vanderbilt.edu⟩.

Received for publication: 12/07/05; accepted for publication: 03/20/06.

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 papers are perfectly distinguishable in terms of methodology.

Very recent research by Bernstam et al.[9] 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, Bernstam et al.[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 Bernstam et al.[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 Bernstam et al.[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 method 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 analyze 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

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 Bernstam et al.[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 to be reviewed and are truly negative and not negative by omission).* 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.

ACPJ-treatment, ACPJ-etiology

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 the ACP Journal.[10]

We used for the present study a modified version of the ACPJ corpus as in Aphinyanaphongs et al.[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.

*In Bernstam et al.,[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 Bernstam et al.[9] In other words, since all the documents are assumed labeled by Bernstam et al.,[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 Bernstam et al.,[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.

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] in August 2005 and August 2002, respectively. These collection dates allowed approximately 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 papers with counts out of 7,810 papers in the gold standard. For the ACPJ gold standard, we obtained 13,279 papers that had counts out of 15,786 papers 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.

Impact Factor

An impact factor of a journal is the average number of citations an article published in this journal receives in 2 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.

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 Aphinyanaphongs et al.[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).

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).

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 0.001.

Performance Metrics

Among the many classifier performance metrics such as precision, recall, average 11-point precision, F1 score, breakeven point, accuracy, error, and area under ROC curve (AUC) that have been used for two-class text categorization (for example, see References 21–25),[21–25] 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, F1 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.† 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 Hand and Till.[27]

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 practice, 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.

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 Scheffer[31] and Dudoit and Van Der Laan.[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 built by using training data exclusively.

†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]

Table 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]

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 predictors 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 analyzing the importance of terms for inclusion in the SSOAB has the following three steps:

1. 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.
2. If the performance of HITON_PC features matched one of the entire feature set (i.e., without feature selection, which is the 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.
3. Finally, we computed contribution Δ_i of each selected feature *i* on step 2 to the SVM model's objective function as described in Guyon et al.[37] We report the normalized contribution of each feature which is equal to $\Delta_i/\Sigma\Delta_i$.

Results

Experiment 1

The results for Experiment 1 are shown in Table 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.

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 2.

These results are interesting since the SSOAB editors are not operating with explicit selection criteria.[11] 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 neo-

Table 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	pancreaticoduodenectoml	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

plasms. 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 Aphinyanaphongs and Aliferis.[35]

Experiment 2

The results of experiment 2 are shown in Table 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]

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 citation count). Content alone compared to content with impact factor showed a statistically significant, but negligible improvement in AUC.

Experiment 3

Table 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).

Experiment 4

Table 5 and Table 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 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

Table 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]

Table 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

SSOAB-based NS model that have AUCs of 0.691, 0.670, 0.673, and 0.772 respectively (see Table 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 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 Bernstam et al.[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 Bernstam et al.,[9] a comparison central to the main hypothesis of the present paper. While we use the gold standard and metrics that Bernstam et al.[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, Bernstam et al.[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 Bernstam et al.:[9] by testing performance on all of MEDLINE, Bernstam et al.[9] do 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, Bernstam et al.[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

Table 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 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]

query and a priori is expected to vary considerably from query to query. Bernstam et al.[9] use 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 Bernstam et al.[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 Bernstam et al.[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 Bernstam et al.[9] to the best of our knowledge). The problem is alleviated to a large extent since Bernstam et al.[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: 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 reviewers may have recommended that a citation needs 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 across 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. This reduces the usefulness

of citation-based metrics 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 highly non-specific but also unavailable when needed the most (i.e., for articles published in recent years).

How feasible and practical is it to build GSS focused filters for identifying high quality articles? Several examples of recent research have 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–for example 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.‡ By applying the SSOAB machine learning model threshold that corresponds to the ROC curve point with sensitivity and specificity of 85% and 85% correspondingly (such points can be obtained for example from the ROC curve with AUC of 0.89 from Experiment 1), 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 to the AUC would improve identification of high quality articles. For example, at 90% sensitivity and 93% specificity (as can be obtained for example from the ROC curve with 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. Furthermore, 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, and so on.

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 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.

‡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.

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;12:207–16.
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;1:447–58.
3. Wilczynski N, Haynes B. Optimal Search Strategies for Detecting Clinically Sounds Prognostic Studies in EMBASE. J Amer Med Inform Assoc. 2005;12:481–5.
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;3:363–69.
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;19:321–7.
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:A-15–A-16.
11. SSOAB. Available at: http://www.surgonc.org. Accessed December 5, 2005.
12. Web Of Science. Available at: http://www.isinet.com/products/citation/wos. Accessed December 5, 2005.
13. Journal Citation Reports. Available at: http://www.isinet.com/products/evaltools/jcr. Accessed December 5, 2005.
14. Salton G, Buckley C. Term weighting approaches in automatic retrieval. Information Processing and Management 1988;24:513–23.
15. Porter M. An algorithm for suffix stripping. Program 1980;14(3):130–7.
16. MEDLINE Stopwords. Available at: http://biolib.princeton.edu/instruct/MedSW.html. Accessed December 5, 2005.
17. Leopold E, Kindermann J. Text Categorization with Support Vector Machines. How to Represent Texts In Input Space? Machine Learning 2002;46:423–44.
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. Available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm. Accessed December 5, 2005.
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 at: http://www.csie.ntu.edu.tw/~cjlin/libsvm. Accessed December 5, 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;45:171–86.
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, Gauvreau K. 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 2003:371–6.
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, (ed). Can citation indexing be automated? Washington, DC: National Bureau of Standards; 1965.
40. PubMed. Available at http://www.ncbi.nlm.nih.gov/PubMed. Accessed December 5, 2005.
41. Jenkins M. Evaluation of Methodological Search Filters—A review. Health Inf Lib J 2004;21:148–63.

Appendix 1

HITS Curves and Precision–Recall Curves from Cross-validation Experiment

The HITS and precision–recall curves were generated from cross-validated experiments. Both curves were generated independently within each testing set of the cross-validation, and an average composite curve for both metrics was generated as an average over the curves from each testing set (Figures 1 and 2).

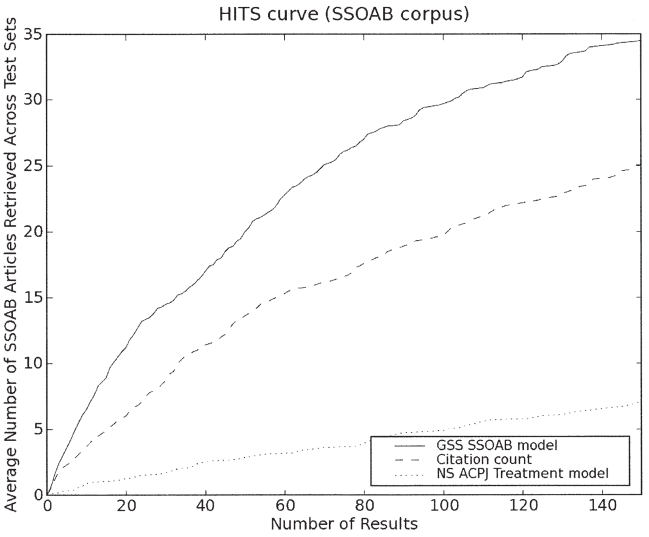

Figure 1. Average HITS curves used on SSOAB corpus. The GSS SSOAB model returns the most true positive documents in the first 150 articles. Citation count and the NS ACPJ Treatment Model applied to the SSOAB corpus return fewer true positive documents in the top 150 returns. The SSOAB corpus was composed of 431 positives and 7,379 negatives.

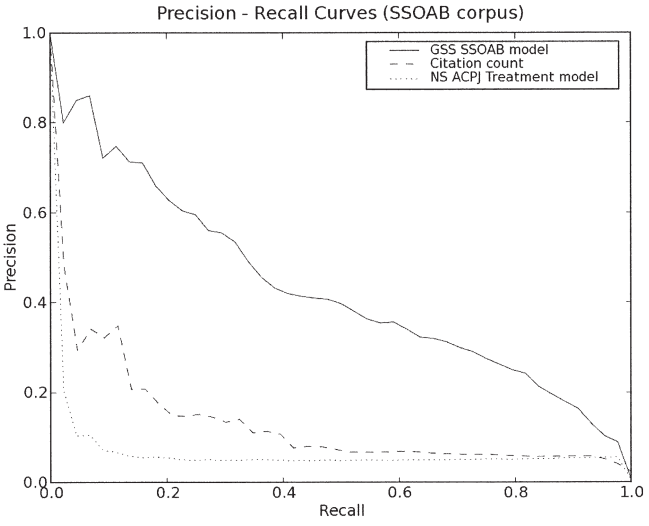

Figure 2. Average precision-recall curves used on SSOAB corpus. The GSS SSOAB model returns the best performing precision-recall curve. Citation count and the NS ACPJ Treatment Model have curves below, thus performing lower than, the GSS SSOAB model. The SSOAB corpus was composed of 431 positives and 7,379 negatives.