IMPROVING BIOMEDICAL INFORMATION RETRIEVAL CITATION METRICS USING MACHINE LEARNING

By

Lawrence D. Fu

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, 2008

Nashville, Tennessee

Approved:

Professor Constantin F. Aliferis

Professor Nunzia B. Giuse

Professor Daniel R. Masys

Professor Cynthia S. Gadd

Professor Lily Wang

ACKNOWLEDGEMENTS

I would like to thank my committee members Cynthia Gadd, Nunzia Giuse, Daniel Masys, and Lily Wang who provided an enormous amount of help over the years. I have truly enjoyed my time working with my advisor Constantin Aliferis who has always impressed me with his dedication and passion for work. Alexander Statnikov and Yindalon Aphinyanaphongs have provided me with crucial help throughout the course of this work. Also, I would like to thank the Department of Biomedical Informatics at Vanderbilt. I realize that I was lucky to have spent my time in such a supportive environment. Finally, I thank my family for their support and patience.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Today, healthcare has become a data driven enterprise involving the analysis and storage of information in large data repositories. This universal need spans from basic biomedical research to clinical care. One type of data is the scientific literature, and the evaluation of the literature is an increasingly integral part of biomedical research and evidence-based medicine [1, 2]. Basic science researchers have used the literature for entity recognition, information extraction, and hypothesis generation [3]. Data mining techniques have also been applied to the literature for drug discovery [4]. On the other hand, clinicians have used the literature to answer questions as part of clinical care [5-11]. Both researchers and clinicians face the daunting task of identifying high quality articles among the existing and growing literature. It is impractical for them to manually monitor the literature, and automated tools have been developed to perform this task. The focus of this work was to improve the performance and usability of existing tools by applying machine learning methods. The thesis consisted of three specific aims:

- Analyze the topic-sensitivity of evaluation methods for journals, articles, and websites
- Examine the feasibility of predicting future citation counts with information available only at the time of publication
- Examine the feasibility of automatically discriminating between instrumental and non-instrumental citations

A Comparison of Evaluation Metrics for Journals, Articles, and Websites in Terms of Sensitivity to Topic

Popular evaluation methods include journal impact factor for journals [12], PubMed's clinical query filters and machine learning-based filter models for articles [13, 14], and PageRank for websites [15]. Previous work has focused on the average performance of these methods without considering topic. This section focused on a subtle but important property: *stability over topics*. It is unknown how performance varies for specific topics or focused searches. A method with excellent average performance may fail in a focused domain, and users should be aware if a method's performance diverges from expected average performance. This section studied the performance of citation metrics (i.e., journal impact factor and PageRank), Boolean queries, and machine learning methods to quantify their variability for different topics.

Machine Learning Models for Predicting and Explaining Citation Count of Biomedical Articles

The most popular method for evaluating the impact and quality of an article is the *citation count* which is the number of citations received by an article within a pre-specified time horizon [16]. A limitation of citation count is that it is unavailable at publication time. This section investigated the feasibility of predicting future citation counts with information available only at the time of publication. The main benefit is improving the usability of citation counts which could accelerate the assessment of research impact and dissemination of new knowledge. Support vector machine (SVM) models were trained on a combination of content-based and bibliometric features. Content features were terms from the title, abstract, and MeSH terms of an article. Bibliometric features included information about the journal or authors. In addition to

the model-building effort, the models were analyzed to identify factors that correlate strongly and potentially determine the chances of an article being cited by many subsequent articles.

Machine Learning Models for Automatic Classification of Instrumental Citations

The use of citation count as an evaluation metric assumes that a citation is an indicator of quality. This is not necessarily true since a citation may serve many purposes unrelated to recognizing the value, rigor, or authority of a cited paper [17-19]. Cited papers may provide background information or acknowledge prior work that influenced the current work. Moreover, citations may serve non-scientific purposes due to social-psychological factors [16, 20, 21]. Thus, a citation is an indirect metric of impact without a single unambiguous use. If instrumental citations can be reliably distinguished from non-essential ones, it may be possible to improve the performance of existing evaluation methods by excluding non-instrumental citations. For the purposes of this work, a citation was operationally defined as instrumental if either of the following were true: the hypothesis of the citing work was motivated by the cited work, or the citing work could not have been completed without the cited work. This section focused on examining the feasibility of automatically classifying citations as instrumental or non-instrumental. A learning approach similar to the one used for predicting citation count was used. SVM models were trained on content and bibliometric features, and performance was evaluated on a manually labeled corpus.

The remainder of the thesis is organized as follows. Chapter II provides a brief review of citation analysis. Chapters III-V present work for each of the three main focuses. Each chapter contains introduction, methods, results, and discussion sections. Chapter VI presents a summary and discussion of the work as a whole.

CHAPTER II

BACKGROUND: REVIEW OF CITATION ANALYSIS

A document may cite another document for a variety of reasons: to acknowledge prior work, identify methodology, or provide background reading. The citing document may be a comprehensive review that attempts to cite the most recent documents on the topic, or the cited article may be highly controversial. On the other hand, a citation may criticize another work and not be an endorsement. Garfield created one of the earliest lists for the many possible reasons for a citation [22]:

1. Paying homage to pioneers
2. Giving credit to related work
3. Identifying methodology, equipment, etc.
4. Providing background reading
5. Correcting one's own work
6. Correcting the work of others
7. Criticizing previous work
8. Substantiating claims
9. Alerting to forthcoming work
10. Providing leads to poorly disseminated, poorly indexed, or uncited work
11. Authenticating data and classes of fact (physical constants, etc.)
12. Identifying original publications in which an idea or concept was discussed

13. Identifying original publication or other work describing an eponymic concept or term
14. Disclaiming work or ideas of others (negative claims)
15. Disputing priority claims of others (negative homage)

Researchers have estimated the frequencies of various types of citations by creating classification schemes. Two approaches for classifying citations are analyzing the articles and interviewing the authors. Article analysis involves examining document text to determine the nature of the relationship between the citing and cited articles. Moravcsik and Murugesan manually reviewed 30 articles in theoretical high energy physics and classified articles in each of the following 5 categorizations [23]:

- Was the article cited for a concept/theory or for a tool/technique?
- Is the cited work necessary for understanding or is it merely an acknowledgment of prior related work?
- Did the cited work provide a foundation for the citing work, or is it an alternative?
- Are the claims of the cited work confirmed or disputed?
- Is the citation essential or redundant?

In their study, most citations were necessary for understanding, provided a foundation for the citing work, and were essential. Chubin and Moitra [24] performed another context analysis study on Physics articles. In their corpus, most citations were categorized as "basic essential" (cited papers were central to the reported research), "subsidiary essential" (not directly related but still essential), or "additional supplementary" (independent supportive observations).

The second approach for identifying the motivation of a citation is interviewing the original author. A questionnaire or personal interview is typically used. Brooks [25] interviewed 20 authors and classified citations into three categories. The first one was persuasiveness, positive credit, currency, and social consensus. The second category was negative credit, and the third category was reader alert and operational information. Brooks found that persuasiveness was the predominant reason for a citation. Cano [26] used Moravcsik and Murugesan's classification scheme to create a questionnaire for authors in structural engineering. In this study, the most popular citation type was a perfunctory citation that acknowledged other work in the same general area as the citing article.

Analyzing the article text and interviewing the authors both have limitations [16]. Article analysis requires much time and effort if it is manually performed. Also, it may be difficult or impossible to identify an author's motivation from the text alone. Author interviews may not reveal the original motivations, and authors may not remember correctly since much time has passed after writing the article. Also, they may not be honest about the purpose of a citation, or there may be inconsistencies between authors [16].

There are other drawbacks to using citation count as a quality metric in addition to the lack of a single unambiguous use of a citation. Citation counts are unavailable at publication since citations accumulate over time. Other problems include the inaccuracy and incompleteness of citation databases as well as the variable citation rates between fields [17-19]. As a result, the validity of citation count as useful metric for the quality of scientific work has been debated.

There are two major theories for explaining the motivations of citations [27]. The normative theory of citing behavior claims that authors cite a paper to indicate that it was an intellectual or cognitive influence on their work [28]. Citations allow scientists to credit colleagues whose work has been useful to them. The theory claims that citations are an indicator of the quality of the cited content and that science is a normative institution built upon internal rewards and sanctions. Merton argued that scientific contributions are evaluated by a set of norms that involve the open communication of ideas, emotional neutrality in the evaluation of one's ideas, and the acknowledgment of intellectual debt to a piece of scholarship [28]. This theory supports the use of bibliometric metrics such as citation count for assessing the quality of research and impact of scientific work.

The social constructivist theory argues that the content of an article has little influence on how it is cited [16]. The theory asserts that scientific knowledge is socially constructed through the manipulation of political and financial resources and that the main purpose of citations is persuasion or rhetoric. Authors can be motivated by factors independent of the quality of the cited work. Examples include defending their claims against attack, advancing their interests, convincing others, or promoting themselves in the scientific community. Supporters of this theory believe that citations cannot accurately measure the quality of papers.

An important question is whether citation count can still provide a useful measure for the impact of a work despite the many motivating factors for a citation. Empirical studies have shown that citations are in fact an informative measure. Cronin [29] reviewed studies that demonstrated the correlation between citation count and other

metrics for the impact of scientific work such as research funding, academic prestige, and peer assessment. Cronin believed that most of the evidence "seems to suggest that scientists typically cite the works of their peers in a normatively guided manner, and that these signs (citations) perform a mutually intelligible communicative function" [29]. White echoed the same sentiment by stating that "results are better explained by Robert K. Merton's norm of universalism, which holds that citers are rewarding use of relevant intellectual property, than by the constructivists' particularism, which holds that citers are trying to persuade through manipulative rhetoric" [30].

CHAPTER III

A COMPARISON OF EVALUATION METRICS FOR JOURNALS, ARTICLES, AND WEBSITES IN TERMS OF SENSITIVITY TO TOPIC

Introduction

The size of the biomedical literature and the web make it difficult to find high-quality documents among the large number of articles, journals, and websites. Automated methods have been developed since manually monitoring the literature and web is impractical. Journal quality is typically measured with impact factor [12]. High-quality articles are identified with PubMed clinical query filters [14] which are a methodological and content criteria-based approach. Machine learning methods such as polynomial support vector machine (SVM) models have been recently introduced as pattern recognition query filters for identifying high-quality articles [13]. The most popular way to rank web pages is PageRank [15].

The methods can be classified as query-independent or query-dependent methods. Query-independent methods are built independently of the learning task and do not consider the query topic. Impact factor, clinical query filters, and PageRank are examples. Impact factor and PageRank are also citation-based methods which are flexible, efficient, and easy to use since they count the number of citations received. However, the flexibility of these approaches can also be a limitation. A document may cite another document for a variety of reasons: to acknowledge prior work, identify methodology, provide background reading, disclaim work of others, or dispute priority

claims [22]. Thus, the citation may not necessarily be an endorsement. For a more thorough discussion of citation analysis, see Chapter II.

Query-dependent methods consider the search topic and are built for the learning task. Two examples are machine learning filter models and the topic-specific impact factor which is presented in this work. Machine learning methods have outperformed citation-based methods in finding high-quality articles [31], and they have several advantages over clinical query filters. They have superior performance, are automatically generated, and allow users to specify a desired sensitivity or specificity. On the other hand, clinical query filters are easier to understand and are more suitable for standard PubMed interfaces.

Previous studies have measured the performance of these methods for all topics, and the variability of these methods for different topics is unknown. It is possible for a method with excellent average performance to fail in a focused domain. Suppose we have a set of articles about two topics (A and B) where 90% of the articles relate to topic A and the remaining articles are about topic B. If a method has a sensitivity of 1 for topic A and .1 for topic B, overall performance would be relatively high. However, a researcher interested only in topic B would unknowingly experience much worse than expected performance.

Topic sensitivity in web-related research is known as topic drift [32] where the highest ranked results are not necessarily related to the query topic. For example, PageRank-based rankings may not yield the best results for a specific topic. Suppose we are interested in topic A and have two web pages with different degrees of relevance to topic A. The first page has a high PageRank, is only marginally relevant to topic A, and

receives most of its links for its discussion of topic B (i.e. most of its links come from pages related to topic B). The second page has a slightly lower overall PageRank, but the majority of its links are related to topic A. PageRank scores the first page higher than the second page although the second page is a better resource for the topic of interest. Topic drift is important since it may lead to sub-optimal results for queries focused on a specific topic or condition in a health-related search. Previous research has discussed topic drift for link analysis algorithms such as PageRank and HITS [32-34]. These approaches rank pages prior to a query and analyze the link structure without considering the topic of a page or the reason for the link. Consequently, high ranking pages are not necessarily related to the query topic. Haveliwala, Richardson, and Nie modified PageRank to consider topic while evaluating webpages [35-37].

The purpose of this work was to determine if performance varies for different topics when evaluating journals, articles, and websites. The specific methods studied were journal impact factor, clinical query filters, machine learning pattern recognition filters, and PageRank. It is possible for a method to perform excellently on average but struggle significantly in a restricted domain. Furthermore, it is unknown how much performance varies for specific topics or focused searches. This issue may affect many clinicians, researchers, and users who are unaware of it.

Methods

This section will present methods for evaluating journal, articles, and websites. For each document type, there will be three sections. First, each method will be explained. Second, the experimental setup and analysis of topic-sensitivity will be discussed. Third, details will be presented on how the corpus or experimental data sets were compiled.

Evaluation Methods for Journals

Journal Impact Factor

The journal impact factor evaluates journal impact regardless of publication size or frequency [12, 38, 39]. It affects journal readership and helps researchers determine to which journal they submit their work. Essentially, it is the average number of citations received per article published in the journal. It is defined for a year y as the quotient of two terms [12]:

$$\text{Impact Factor} = \frac{\text{Number of citations in year } y \text{ to journal items published in years } (y-1) \text{ and } (y-2)}{\text{Number of journal articles published in years } (y-1) \text{ and } (y-2)} \quad (1)$$

The numerator is the number of citations received in a given year to journal items published in the previous two years. The denominator is the number of journal articles from the previous two years. Items in the numerator include articles, editorials, and letters to the editor, while the denominator consists only of articles [12]. For example, the impact factor of the New England Journal of Medicine (NEJM) for 2004 is the

number of citations in 2004 to its published items from 2002 and 2003 divided by the number of articles from 2002 and 2003.

Topic-Specific Impact Factor

Prior work considered the impact factor of topics irrespective of journal by computing the number of citations received by articles in a topic area (e.g. asbestos) [40, 41]. However, this metric does not assess journals. A formula is needed to isolate the contribution of a specific topic from the overall impact factor to study the sensitivity of impact factor to topic. A topic-specific impact factor (TIF) can be calculated for a journal in year *y* by considering only publications related to a given topic:

$$\text{TIF} = \frac{\text{Number of citations in year } y \text{ to items published in years } (y-1) \text{ and } (y-2) \text{ that were relevant to topic}}{\text{Number of journal articles published in years } (y-1) \text{ and } (y-2) \text{ that were relevant to topic}} \quad (2)$$

For example, the numerator of the cardiology-specific impact factor of NEJM in 2004 is the number of citations in 2004 to cardiology-related items published in NEJM from 2002 and 2003. The denominator is the number of cardiology-related articles. Determining topic relevance is topic-specific. For example, we can consider an item relevant to cardiology if its MEDLINE record contains the MeSH term “Cardiology”, a related topic such as “Cardiovascular Diseases” that is specified in the “See Also” field of the MeSH record, or a term residing in a sub-tree of these terms [42]. When specifying topics, the topics do not need to be exclusive or cover all items for the adjustment to be meaningful.

Topic-Mix Adjusted Impact Factor

Impact factor can be adjusted for a mix of topics by computing a weighted average of the topic-specific impact factors. The topic-mix adjusted impact factor for *k* topics can be defined as:

$$\text{Topic-mix adjusted impact factor} = \sum_{i=1}^{k} w_i \times TIF_i \quad (3)$$

TIF_i is the topic-specific impact factor of topic *i*, and w_i is a weight proportional to the importance of topic *i* normalized such that the sum of all weights equals one and each weight is between 0 and 1. For example, a researcher interested in gastroenterology twice as much as hematology would weight the topic-specific impact factors of gastroenterology and hematology by 2/3 and 1/3 respectively. If all topics are weighted equally, the topic-mix adjusted impact factor is the arithmetic mean of the topic-specific impact factors for all topics.

Analysis for Journal Methods

When computing topic-specific impact factors, there were no p-values or confidence intervals since they were population totals and not point estimates. Variability was analyzed by calculating the absolute difference of impact factor to topic-specific impact factor. The minimum, median, maximum, and interquartile ranges of the differences were computed to assess the skewness and spread of the values. Interquartile range measures dispersion and is the difference of the third and first quartiles. There

should be little discrepancy between the methods if citations are evenly distributed over topics.

A Bland-Altman plot [43] was used to determine whether topic-specific impact factors coincide with impact factors or if they are significantly different. This plot shows whether a new measurement method agrees with another method by plotting the measurement differences against their mean and illustrating any dependence between the values. The correlation coefficient was considered but was determined to be an inappropriate method since it can be misleading [43].

Another consideration was whether variation was randomly caused by smaller sample sizes independently of topic. By definition, journal impact factor is calculated on a larger number of publications than the topic-specific impact factor. To determine whether the difference between the two measures is associated with sample size, regression coefficients were computed for the following regression model:

$$\text{Diff(TIF,IF)} = \beta_0+\beta_1*(\text{sample size difference}) + \beta_2*\text{topic} + \beta_3*\text{year} + \beta_4*\text{journal} \quad (4)$$

where Diff(TIF, IF) is the difference between topic-specific impact factor and impact factor, and "sample size difference" is the difference between the number of articles used in each calculation. The "topic", "year", and "journal" variables are categorical variables representing different values for the topic, year, and journal. They were included in the model to account for any possible confounding effects.

Experimental Data Set for Journal Methods

A data set was created by identifying all articles for a set of journals, topics, and time periods. Six journals were chosen: Annals of Internal Medicine (AIM), American Journal of Medicine (AJM), British Medical Journal (BMJ), Journal of the American Medical Association (JAMA), Lancet, and New England Journal of Medicine (NEJM). These journals were selected since they are a collection of well-known journals with a wide range of impact factors. Eight general topics of internal medicine were used along with a randomly selected set of narrowly-defined subtopics. The topics were defined by the MeSH vocabulary. The eight topics were Cardiology, Endocrinology, Gastroenterology, Hematology, Medical Oncology, Nephrology, Pulmonary Disease, and Rheumatology. The narrowly-defined subtopics were Esophageal Diseases, Gastroenteritis, Gastrointestinal Neoplasms, Hernia, Intestinal Diseases, and Stomach Diseases. For each journal and topic, all relevant MEDLINE records were retrieved for 2003 and 2004, and citation counts and journal impact factors were obtained from the ISI Web of Knowledge [44].

Evaluation Methods for Articles

Clinical Query Filters

Clinical query filters were originally designed by Haynes and colleagues [14] and are the most widely available method for identifying high-quality articles through PubMed. These filters are semi-manually constructed Boolean queries of MeSH terms, publication type, or text word fields of the MEDLINE record. All articles that match a

given combination of terms are returned. Performance is measured by sensitivity and specificity. Filters are defined for diagnosis, etiology, prognosis, and treatment with queries optimized for sensitivity and specificity [45]. For example, the specificity-optimized filter for therapy is: (randomized controlled trial [Publication Type] OR (randomized [Title/Abstract] AND controlled [Title/Abstract] AND trial [Title/Abstract])). This query returns all articles with publication type "randomized controlled trial" or with all three words in the title or abstract.

Support Vector Machine Models

Machine learning methods provide another approach to identifying high-quality articles. In previous research, polynomial support vector machine models had superior performance compared to the clinical query filters [13]. These models preprocess fields and text from MEDLINE records for use as features during learning. A kernel function maps the input space to a "feature" space where a hyperplane is calculated to separate the classes of data. The models learned from a previous study were used [13]. Performance was measured by area under the receiver operating curve (AUC).

Analysis for Article Methods

Absolute differences were computed between the performance when ignoring topic (i.e., all articles included) and the performance for a subset of articles related to a given topic. Evaluation metrics were sensitivity and specificity for clinical query filters, and the metric was AUC for the SVM models. The minimum, median, maximum, and interquartile ranges of these differences were also computed. Second, Wilcoxon signed

rank tests were performed which test the difference between paired measurements. They compare repeated measurements after an experimental manipulation to determine if a value has changed. The null hypothesis is that the difference is zero. A p-value less than .05 means that the difference is significantly different from zero, which implies that the method does not retain performance for individual topics.

Experimental Data Set for Article Methods

The experimental corpus was the same as the corpus used in previous work to compare clinical query filters and SVM models [13]. The gold standard was the ACP Journal Club [46]. It is a meta-publication where experts review internal medicine journals on a monthly basis to identify high-quality articles for categories such as diagnosis, etiology, prognosis, and treatment. All MEDLINE articles from the ACP Journal club during the study period were positive cases or considered high-quality. The remaining articles from the journals during the same period were negative cases or not considered high-quality. For the treatment and etiology categories, there were 15,786 MEDLINE records from July 1998 to August 1999. For prognosis and diagnosis, there were 34,938 MEDLINE records from July 1998 to August 2000. The longer timeline enabled a sufficient number of positive cases. Articles were converted into a format suitable for the learning methods by extracting and encoding terms from the abstract, title, MeSH terms, and publication type.

A set of 18 MeSH terms was randomly selected to cover a range of topics. The topics were: Bone Diseases, Cardiovascular Diseases, Cysts, Diabetes Mellitus, Endocrine System Diseases, Gastroenteritis, Gastrointestinal Diseases, Heart Diseases,

Hematologic Diseases, Hernia, Infection, Kidney Diseases, Lung Neoplasms, Myocardial Infarction, Muscular Diseases, Neoplasms, Respiratory Tract Diseases, and Rheumatic Diseases. Articles were relevant to a topic if its MEDLINE record contained the MeSH term or a term residing in a sub-tree.

Evaluation Method for Websites

PageRank

PageRank is a citation-based method for evaluating the quality of web pages [15]. It is motivated by the intuition that high quality pages will link to other high quality pages. Specifically, it is calculated by modeling user behavior as a random surfer ignoring page content by either following a link arbitrarily or jumping randomly to another page. The PageRank of a page is proportional to the likelihood that the surfer will visit it. The PageRank of a page u is calculated as follows:

$$PR(u) = \frac{1-\alpha}{N} + \alpha \sum_{v \in B_u} \frac{PR(v)}{|F_v|}$$

where N is the total number of web pages in the network, B_u is the set of pages linking to page u, and F_v represents the set of pages to which page v links. The term α is a parameter specifying the probability of following a link or randomly jumping to a page. The surfer will jump to a random page with probability 1 - α and follow an outlink with probability α. It is usually 0.85 but can be any value between 0 and 1. The first term of the equation is the probability of randomly jumping to a given page. The second term is

the sum of all PageRanks for its incoming links. For each inlink, the PageRank of the original page is divided by the number of outlinks for that page. These values are summed over all incoming links and weighted by α. A vector of PageRank values is defined over all pages in the network, and each page is initialized with an equal value. PageRank calculations are performed iteratively as matrix operations until the PageRank values converge, which is guaranteed by adding links to pages without any links and having the random jumps.

There have been of number of modifications to PageRank to address topic drift. Haveliwala computed topic-sensitive PageRank scores by calculating a score for each page with respect to a number of topics [35]. The topics were top level categories from the Open Directory Project. The topic-sensitive PageRanks were computed by biasing the random jump to favor pages related to a given topic, and the final PageRank values were computed at query time by weighting each topic-sensitive PageRank according to how similar a topic was to the query. Richardson [37] used an intelligent surfer model to analyze the content of a webpage. The probability of following a link or jumping to a page was proportional to the relevance of a page to the query. Nie [36] augmented the random surfer model by using a topical random surfer that considered topics while surfing. When a surfer follows an outlink, it can stay on the same topic or change the topic of interest.

Analysis for Websites

Experimental Considerations for Websites

Web-related research is challenging since it is difficult to replicate real-world conditions. The size of the web makes experiments computationally intensive, and web crawlers cannot determine if they have detected all incoming links to a page. Researchers typically sample pages to create a static snapshot of the web. PageRank values are affected when pages are removed during sampling since the network topology changes as links are removed. It is not completely understood how sampling affects the stability of rankings [33, 47, 48].

Studying the topic-sensitivity of PageRank required understanding the ramifications of sampling to ensure that any observed variability was not caused by sampling. The first consideration was sampling networks by selecting pages from the same domain. Kamvar demonstrated that most pages link to pages from the same domain [49]. He found that 83.9% of links connected pages from the same domain in the January 2001 Stanford WebBase crawl. WebBase is a collection of crawls of websites used for web research. The percentage rose to 95.2% after pages without outlinks were removed. Sampling pages from the same domain appears to minimize the effect on PageRank.

Another sampling criterion was to select high-ranking pages. Ng [48] showed that removing pages with low PageRank did not affect the stability of the top 10 results. My study investigated whether rankings are stable for all results since users may be interested in more than 10 results. Four domains were chosen from WebBase [50]: the National Diabetes Education Program (NDEP), the National Eye Institute (NEI), the

National Heart Lung and Blood Institute (NHLBI), and the Centers for Disease Control and Prevention (CDC). These domains were selected to provide biomedically relevant samples of various sizes.

PageRanks were first computed for all pages within a domain. Then the pages with the lowest PageRanks were removed, and PageRanks were computed for the remaining pages. The stability of the rankings was measured with Haveliwala's Ksim metric [35], which is based on Kendall's τ distance measure. Ksim is the fraction of pairwise ranking comparisons that are consistent between two rankings lists. If page A is ranked higher than page B in one ranking set, Ksim checks if page A is ranked higher than page B in the other set. For example, a Ksim value of .9 means that 90% of the pairwise comparisons are consistent in both rankings. The two steps of removing pages and calculating PageRanks were repeated until no pages remained. The number of pages removed per iteration depended on the original number of pages in the domain because of computational limitations. Running time became prohibitive for a large number of pairwise comparisons. For the NHLBI and CDC sites, the starting set consisted of the 2000 highest ranked pages, and 100 pages were removed per iteration. For the NEI site, all pages were included, and 10 pages were removed per iteration. For the NDEP site, all pages were included, and pages were removed individually.

Studying the Topic-sensitivity of PageRank

The variability of PageRank for different topics was assessed by removing pages unrelated to a given topic. An initial network included a mixture of topics. If highly-ranked pages in the original network received many links from pages unrelated to a given

topic, then they could decrease in rank within a topic-specific network. First, PageRanks were computed for all pages. Then, pages unrelated to a specific topic were removed, PageRanks were re-computed on this subset, and the similarity between the two rankings sets was measured with Ksim.

As mentioned previously, removing links affects the network topology and PageRank values. Sampling topic-specific networks can alter rankings due to the topic or random fluctuations from the changing topology. The effect of random fluctuations was first determined as a baseline for comparison. Random subsets were generated with the same number of pages as the topic-specific subset and PageRanks were computed for the random subset. Then, the similarity between the original and random subsets was measured with Ksim. If the similarities in rankings for topic-specific subsets were higher than the random subset values, the increase was attributed to topic rather than random changes in the network connectivity.

Two health-related domains were chosen: the Centers for Disease Control and Prevention (CDC) and the National Cancer Institute (NCI). A number of well-represented topics were selected for each domain. For the CDC site, these topics were Genomics, National Center on Birth Defects and Developmental Disabilities (NCBDDD), National Center for Infectious Diseases (NCIDOD), National Immunization Program (NIP), and Tobacco. For the NCI site, the topics were Breast, Cervix, Colon, Lung, and Prostate. Topic relevance was determined by searching the website address for the topic or a related word. For example, a NCI page was included in the "lung" topic if "lung" or "pulm" was in the address. The CDC site was organized in a directory structure based on the topics.

Experimental Data Set for PageRank

The September 2006 crawl from the Stanford WebBase [50] was the source of web pages for this study. The general crawl contained about 90 million pages. WebBase provided link and html information, but only the link structure was needed for the study. The similarity between two sets of rankings was measured with Haveliwala's Ksim metric [35].

Results

Variability of Journal Ranking for Different Topics

The variability of impact factor over topics and journals was measured by adjusting it for topic with equation (2) from the Methods section. Table 1 and Table 2 show that rankings based on impact factor and topic-specific impact factor were not equivalent. A higher impact journal did not always have a higher topic-specific impact factor for a given topic. For example, NEJM had a higher impact factor than JAMA but had a lower cardiology-specific impact factor. There were 10 reversals (8.33% of the comparisons, 95% confidence interval 3.39% to 13.28%) for the 120 comparisons among the 15 journal pairs and 8 topics. There were 3 extreme cases where a journal impact factor was 1.5 times greater than another journal while the other journal's topic-specific impact factor was 1.5 times greater. The topics were nephrology (AJM, BMJ), gastroenterology (NEJM, JAMA), and rheumatology (Lancet, NEJM).

Table 1: Journal impact factor and topic-specific impact factors for general topics in 2004 and 2003

	Journal	Topic-specific Impact Factors for General Topics				Impact Factor
		Cardiology	Endocrinology	Gastroenterology	Hematology	
2004	AIM	16.07	13.85	16.92	7.94	13.11
	AJM	4.09	3.44	2.73	6.38	4.18
	BMJ	7.55	6.48	7.37	5.73	7.04
	JAMA	42.18	28.27	60.55	13.87	24.83
	Lancet	33.8	47.7	18.86	11.98	21.71
	NEJM	37.46	54.31	37.68	33.71	38.57
2003	AIM	14.37	19.83	12.73	10.63	12.43
	AJM	4.21	5.82	2.43	4.3	4.4
	BMJ	7.95	6.84	4.98	5.57	7.21
	JAMA	38.12	28.24	70	13.38	21.46
	Lancet	24.42	34.33	17.91	8.34	18.32
	NEJM	38.05	55.78	33.66	28.78	34.84

Table 2: Journal impact factor and topic-specific impact factors for general topics in 2004 and 2003

	Journal	Topic-specific Impact Factors for General Topics				Impact Factor
		Medical Oncology	Nephrology	Pulmonary Disease	Rheumatology	
2004	AIM	12.49	23.17	12.66	15.4	13.11
	AJM	3.95	4.31	3.1	6.29	4.18
	BMJ	5.57	2.37	7.94	8.77	7.04
	JAMA	35.58	20.32	36.47	13.4	24.83
	Lancet	23.16	14.3	27.41	52.5	21.71
	NEJM	44.8	27.93	37.97	24.08	38.57
2003	AIM	12.14	23.06	13.21	14.5	12.43
	AJM	3.98	5.33	3.44	5.82	4.4
	BMJ	5.76	4.00	5.37	12.25	7.21
	JAMA	39.27	18.94	30.13	12.8	21.46
	Lancet	17.78	14.61	14.12	17.94	18.32
	NEJM	40.46	39.51	22.42	45.33	34.84

Table 3: The minimum, median, maximum, and interquartile ranges for the absolute differences between impact factor and topic-specific impact factor in 2004

Topic	Min.	Median	Max.	IQR
Cardiology	0.09	2.04	17.35	11.58
Endocrinology	0.56	2.09	25.99	15
Gastroenterology	0.33	2.15	35.72	2.92
Hematology	1.31	5.02	10.96	7.53
Medical Oncology	0.23	1.46	10.75	5.61
Nephrology	0.13	6.04	10.64	5.55
Pulmonary Disease	0.45	0.99	11.64	5.1
Rheumatology	1.73	6.86	30.79	12.38

The absolute differences between the two measures and the Bland-Altman plot also indicate that the methods are not equivalent. Table 3 shows the minimum, median, maximum, and interquartile ranges of these differences for all journals and topics. The values were unstable since the maximum differences ranged from about 10 to 35. In Figure 1, the Bland-Altman plot showed that the difference in impact factor and topic-specific impact factor depended on their values, and the divergence increased as the values increased. Also, the difference did not depend on

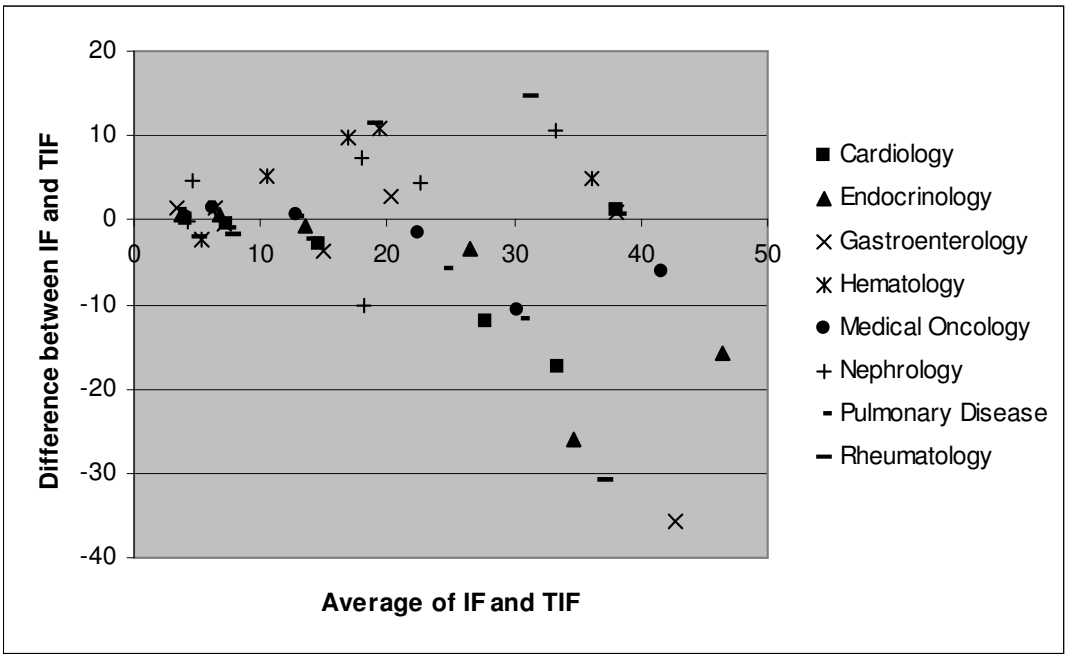

Figure 1: Bland-Altman plot for the differences between Impact Factor and Topic-Specific Impact Factor

specialty since all topics showed some difference. If the methods were in agreement, all values would lie between horizontal lines at -22.17 and 17.7, which is the range of two standard deviations from the mean difference of -2.24. Three values fall outside this range.

The observations for the eight general topics from internal medicine were also evident for gastroenterology subtopics as shown in Table 4. There were a number of ranking reversals as with the general topics. The variation increased for more specialized topics and was most pronounced in the 3 highest impact journals. JAMA had the greatest variability with a maximum topic-specific impact factor that was over 13 times larger than its minimum. For increasingly specialized topics, the overall impact factor became less meaningful. In 2004, JAMA had an impact factor of 24.83, gastroenterology-specific impact factor of 60.55, and topic-specific impact factors for gastroenterology-based subtopics ranging from 6 to 80.07. These results suggest that researchers studying a specific disease should not rely on overall impact factor for journal evaluation.

Table 4: Topic-specific impact factors for the narrowly defined topics, journal impact factor, and topic-specific impact factor for Gastroenterology in 2004. Empty entries had less than 5 articles.

Journal	Topic-specific Impact Factors for Narrowly Defined Topics						Impact Factor	TIF for Gastro.
	Esophageal Diseases	Gastro-enteritis	Gastrointestinal Neoplasms	Hernia	Intestinal Diseases	Stomach Disease		
AIM	-	20.00	17.64	-	18.00	15.25	13.11	16.92
AJM	1.94	7.80	2.82	-	4.82	-	4.18	2.73
BMJ	5.43	4.50	5.90	7.00	6.90	2.33	7.04	7.37
JAMA	9.00	20.00	79.43	-	80.07	6.00	24.83	60.55
Lancet	15.86	28.08	21.26	2.50	19.95	18.60	21.71	18.86
NEJM	20.33	29.83	60.21	7.50	37.48	44.00	38.57	37.68

Additional experiments were performed to ensure that variation was not a random occurrence unique to a single year. First, the experiments were replicated for 2003 and yielded consistent results as shown in Table 1 and Table 2. Many of the relative rankings for journals were retained, and some of the same reversals existed. Also, ranges of topic-specific impact factors were comparable. Next, the regression model in equation (4) verified that variation was not randomly caused by sampling. The regression coefficient for sample size difference, $\beta 1$, was .0021 and not significantly different from zero (p-value = .6062). The difference between topic-specific impact factor and impact factor did not appear associated with differences in sample size.

Data from 2004 was used to provide an example of a topic-mix adjusted impact factor with a topic mix where cardiology was weighted three times more than pulmonary disease. JAMA had a topic-mix adjusted impact factor of 40.75 while NEJM was 37.59. In this case, JAMA had a higher cardiology-specific impact factor, while NEJM had a higher pulmonary disease-specific impact factor. Due to the emphasis on cardiology in this example, JAMA had a higher topic-mix adjusted impact factor despite the fact that NEJM had a higher overall impact factor. This example shows that the unadjusted impact factor may not be the best guide in evaluating journals for topic mixes.

Table 5: The minimum, median, maximum, and interquartile ranges for the absolute differences between overall and topic-specific sensitivity/specificity.

Optimized for	Category	Sensitivity				Specificity			
		Min	Median	Max	IQR	Min	Median	Max	IQR
Sensitivity	Diagnosis	0.02	0.02	0.15	0.0013	0.015	0.087	0.23	0.097
	Etiology	0.028	0.07	0.07	0	0.00047	0.059	0.22	0.10
	Prognosis	0.031	0.1	0.57	0.15	0.0029	0.053	0.18	0.042
	Treatment	0.0035	0.01	0.026	0.0025	0.0027	0.030	0.17	0.053
Specificity	Diagnosis	-	-	-	-	-	-	-	-
	Etiology	0.16	0.34	0.49	0.28	0.0066	0.13	0.31	0.086
	Prognosis	0.11	0.24	0.52	0.33	0.030	0.099	0.22	0.035
	Treatment	0.034	0.053	0.07	0.023	0.00037	0.048	0.13	0.033

Variability of Article Evaluation Methods for Different Topics

Performance of the clinical query filters differed for specific topics. Table 5 summarizes the differences between the overall sensitivity/specificity and the observed values. There was considerable variability for some categories. For example, the sensitivity-optimized prognosis filter had a median difference of 0.1, maximum difference of 0.57, and an interquartile range of 0.15 for sensitivity. These values are relatively large since sensitivity ranges from 0 to 1. The Wilcoxon signed rank tests suggested that performance was unstable for most categories. The p-values were less than .05 for all cases except sensitivity with the sensitivity-optimized diagnosis filter and both values for the sensitivity-optimized prognosis filter.

The SVM models were more stable over topics as shown in Table 6. The AUC values cannot be compared directly with the Haynes' filters results since they are not sensitivity or specificity values. However, AUC values also range from 0 to 1. Differences were much smaller since the largest interquartile range is .065, and the

largest maximum difference was 0.13. The Wilcoxon tests for the SVM models showed that all categories except for diagnosis did not differ significantly from the overall AUC values. These results imply that the SVM models are less sensitive to topic and more stable for specific topics. One important observation for the diagnosis category is that it had few positive documents. A number of the topics had no positive documents, and most of the topics had fewer than 4 positive cases out of several hundred or thousand articles. The diagnosis results may be consistent with the results for the other categories with more positive cases.

Table 6: The minimum, median, maximum, and interquartile ranges for the absolute differences between AUC values

Category	Minimum	Median	Maximum	IQR
Diagnosis	0.0083	0.038	0.04	0.012
Etiology	0.0027	0.028	0.13	0.05
Prognosis	0.0041	0.045	0.10	0.065
Treatment	0.00054	0.0040	0.041	0.0078

Sampling Web Pages with High PageRank Values

Before analyzing the topic-sensitivity of PageRank, it was verified that sampling websites with high PageRank values maintained stable rankings in the remaining pages. Subsets were created by repeatedly removing pages with low PageRanks and re-computing PageRanks. Figures 2-5 show that rankings did not fluctuate dramatically when low ranking pages were removed. All domains had Ksim values over 0.8 after the first removal. The values gradually decreased with fewer pages until a small number remained. The results indicated that samples of high-ranking pages yield subsets with relatively stable rankings.

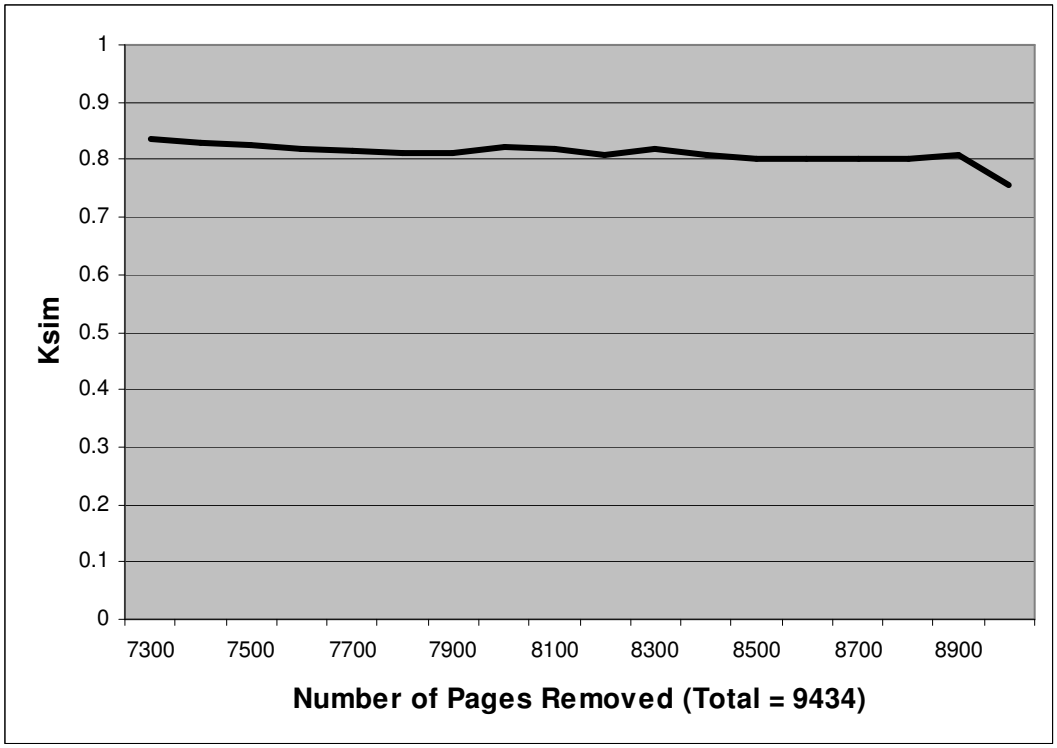

Figure 2: Similarity in rankings for the CDC domain as pages were removed

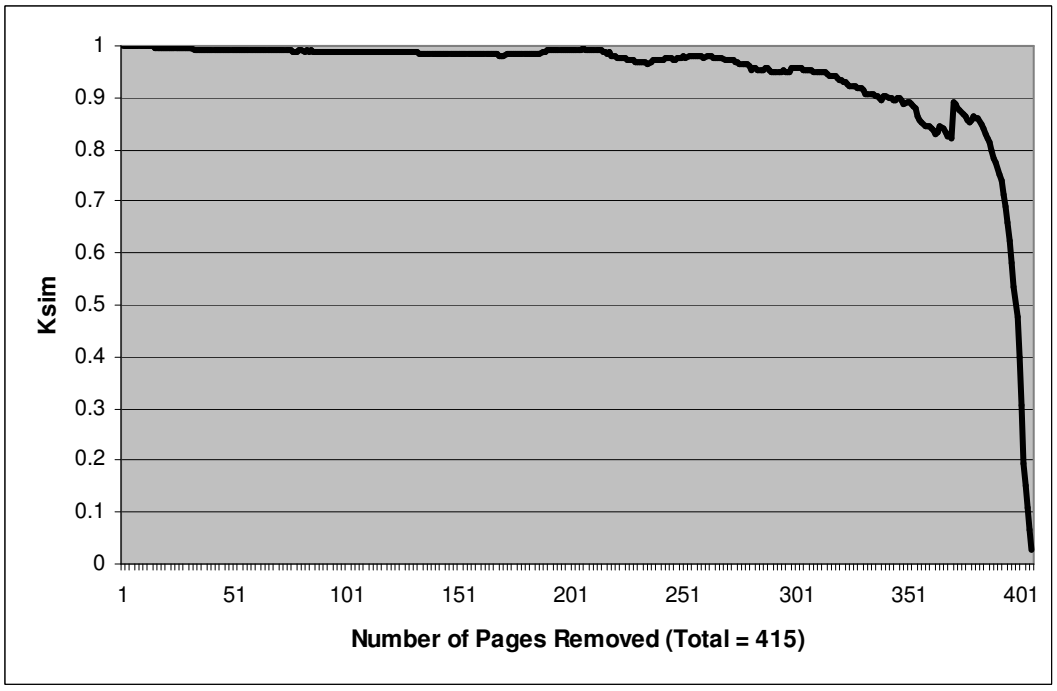

Figure 3: Similarity in rankings for the NDEP domain as pages were removed

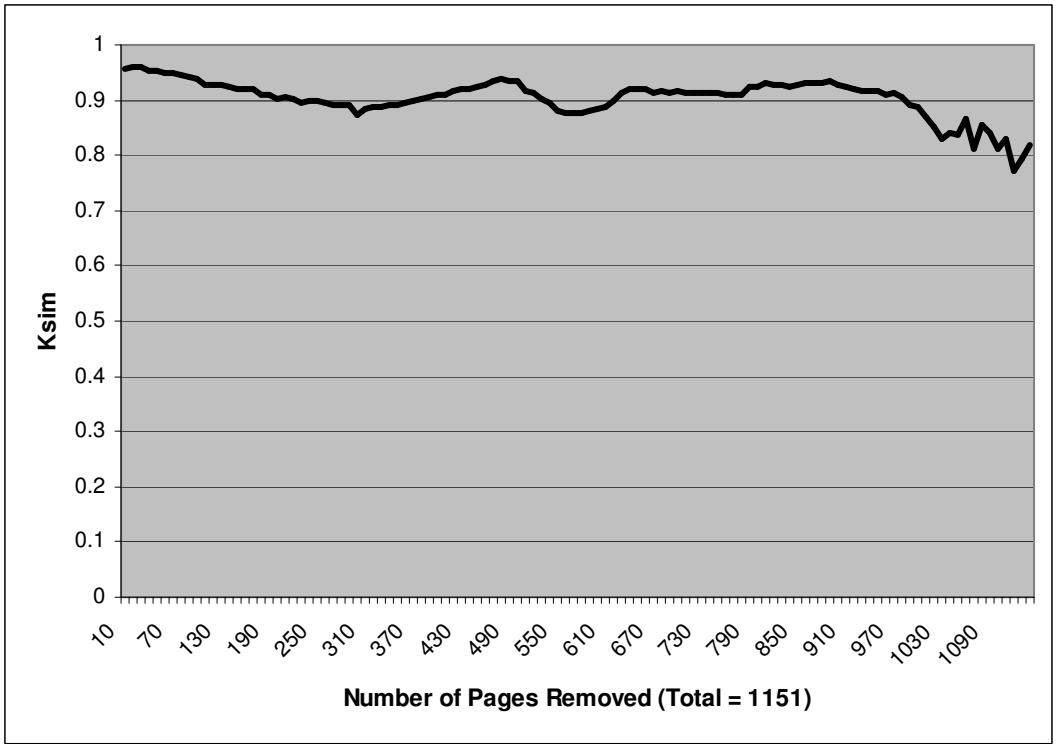

Figure 4: Similarity in rankings for the NEI domain as pages were removed

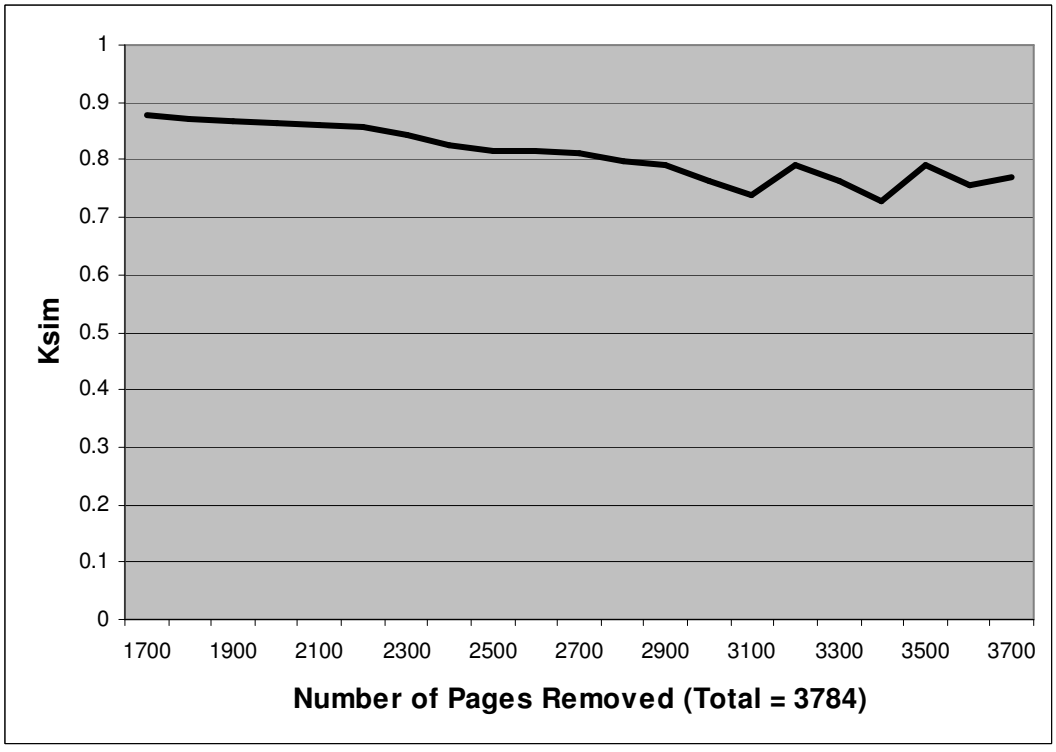

Figure 5: Similarity in rankings for the NHLBI domain as pages were removed

Variability of PageRank for Different Topics

The topic-sensitivity of PageRank was measured by removing pages unrelated to a topic and measuring the stability of rankings. Random subsets were generated to estimate the effect of sampling on rankings. Table 7 displays the Ksim values for all topics. For both domains, Ksim values for the topic subsets were larger than the values for the random subsets. The larger values imply that the rankings were dependent on topic and varied over topics. The CDC topics had higher Ksim values than the NCI topics which meant there was less variability in the CDC rankings. Removing unrelated pages for each topic affected the CDC rankings less than the NCI rankings. This result is explained by the prevalence of intra-topic linking in the original network. Removing unrelated pages does not drastically affect the topology or rankings if pages link mostly to pages within the same topic. Removing unrelated pages will greatly influence the link structure and rankings if most links connect to pages outside the topic.

Table 7: Similarity of rankings for topic-isolated and random subsets along with the percentage of links remaining after topic isolation

Domain	Topic	Number of Pages	Ksim for Topic subset	Ksim for Random subset	Fraction of links within same topic
CDC	Genomics	647	0.97	0.58	0.85
	NCBDDD	725	0.87	0.63	0.71
	NCIDOD	1185	0.79	0.68	0.76
	NIP	357	0.87	0.49	0.83
	Tobacco	482	0.94	0.53	0.87
NCI	Breast	219	0.71	0.31	0.32
	Cervix	204	0.74	0.24	0.42
	Colon	199	0.72	0.20	0.37
	Lung	254	0.76	0.32	0.36
	Prostate	151	0.70	0.24	0.32

Table 7 shows that the Tobacco and Genomics topics in the CDC site had the greatest percentage of intra-topic links (.87 and .85 respectively) as well as the highest Ksim values (.94 and .97). The Breast and Prostate topics in the Cancer site had the lowest percentage of intra-topic links and two of the lowest Ksim values. Rankings were more stable with a greater proportion of links to related pages, and rankings were more unstable with links to unrelated pages.

Discussion

This work studied the variability of evaluation metrics for the scientific literature and web when considering different topics. Previous research studied the average performance of impact factor, clinical query filters, and SVM-based models over all topics. The present study builds on prior work by analyzing the stability of these methods for specific topics. Experimental results demonstrated that impact factor, clinical query filters, and PageRank are sensitive to topic and vary widely for different subjects. Researchers should realize that average performance cannot always be expected for focused searches. Approaches that adjust for topic or are insensitive to topic should be used if available. The topic-specific impact factor and SVM models are two viable options.

Two aspects of PageRank's behavior for evaluating web pages were also investigated as part of this study. First, it was demonstrated that removing pages with low PageRanks is a reasonable sampling method. Experiments showed that the rankings of the remaining pages were relatively stable. Prior work had focused on the stability of

the top 10 results. Second, rankings based on PageRank vary depending on the proportion of links from unrelated pages.

Impact factor, clinical query filters, and PageRank are unstable over topics since they are query-independent methods built separately from the learning task. Citation-based metrics, including impact factor and PageRank, suffer since a citation is not necessarily an endorsement related to a topic of interest. An article may cite another article for many reasons. Even if the citation is an endorsement, the reason for the citation may not be relevant to the query topic and distort topic-specific rankings.

The topic-sensitivity of clinical query filters may be due to their manual creation. Experts choose terms that reflect their expertise. The coverage of terms may not be exhaustive since research areas use different jargon and vocabulary, and some topics may lack adequate consideration. On the other hand, SVM models automatically learn terms for all topics from the corpus, and the machine learning methods should not perform poorly for topics included in the corpus.

This work's findings have practical implications. Relying on topic-sensitive methods can provide misleading conclusions. For example, researchers interested in gastrointestinal diseases would believe that NEJM is the best journal according to impact factor. However, JAMA may be a better choice since it has higher topic-specific impact factors for gastroenterology and gastrointestinal diseases. The variability could result in queries of lower than expected sensitivity or specificity. Similarly, someone interested in finding websites about gastrointestinal diseases could receive less than optimal results if PageRank is used for ranking. The top results may be prominently ranked for their coverage of other topics. Taken together, these consequences represent the potential for a

habitually flawed evaluation of the literature and web. Researchers' work may not be reaching as large an audience as possible, and articles and web sites may frequently be misidentified with respect to quality.

The topic-sensitivity of the methods was not as extensive as the hypothetical examples in the introduction. However, it is still present and should be considered when using the studied methods. For articles and journals, this work was the first step in characterizing the application of these approaches for specific domains. Aphinyanaphongs and colleagues previously showed that it is naive to believe that citation-based metrics can describe all clinical uses [31]. Similarly, this work shows that it is unrealistic to expect impact factor, clinical query filters, and PageRank to exhibit average performance for all clinical contexts. The results support the use of specialized learning methods for focused searches on a given topic. Although this approach can be computationally intensive, there has been evidence that methods designed specifically for a given query or learning task can outperform non-specific or query-independent methods in finding high-quality articles in the literature [31].

CHAPTER IV

MACHINE LEARNING MODELS FOR PREDICTING AND EXPLAINING CITATION COUNT OF BIOMEDICAL ARTICLES

Introduction

The most popular method for evaluating the impact and quality of an article is the *citation count* which is the number of citations received by an article within a pre-specified time horizon [16]. One limitation of citation count is its unavailability before this horizon expires (typically several years after publication). This delay renders citation counts primarily useful for historical assessment of the scientific contribution and impact of papers. Other problems include the inaccuracy and incompleteness of citation databases, variable citation rates between fields, and multiple purposes of citations unrelated to quality [17-19]. For a more complete discussion of citation analysis in general, see Chapter II.

Automatic prediction of citation counts could provide a powerful new method for evaluating articles while alleviating many difficulties associated with the explosive growth of the biomedical literature. Faster identification of promising articles could accelerate research and dissemination of new knowledge. Accurate models for citation count prediction could also improve our understanding of the factors that influence citations.

Predicting and understanding article citation counts is however a challenging problem both on theoretical grounds and on the basis of several decades of related

empirical work. In fact, the bulk of the literature concerning citation counts addresses the motivating factors for article citations rather than predicting them [16].

From a theoretical point of view, it has been found that citation prediction is difficult because of the nature and dynamics of citations [51, 52]. Predictions based on current data assume that citation behavior will not change in the future, and this assumption may be violated in fast-paced research fields such as biomedicine. Citations are a noisy, indirect quality measure, and accumulation rates vary unpredictably between articles. Breakthrough papers can stop receiving citations after review articles replace them or the subject matter becomes common knowledge [51]. Redner identified four major categories for citation behavior [53]. "Sleeping beauties" are highly-cited articles receiving most of their citations long after publication, "major discovery papers" demonstrate a spike of citations after its contribution is recognized, "classic publications" are cited over long periods of time, and "hot papers" increase their citation rate over time. Another difficulty in making accurate predictions is the sparseness of a citation network [52]. Fitting a reliable statistical model is difficult since the number of links is small compared to the number of nodes, and negative cases (i.e., non-connected nodes) grow much more rapidly than positive cases (i.e., connected nodes) [54]. Another contributing factor is that citation rates may have a degree of randomness. For example, a high-impact journal paper may increase the citation rate of papers within the same issue [55].

Previous empirical research predicted long-term citation counts from citations accumulated shortly after publication. In the Knowledge Discovery and Data (KDD) Mining Cup competition of 2003 [56], researchers predicted the evolution of the number of citations received by a set of 441 articles in high-energy physics from arXiv.org during

successive three month periods. arXiv.org is a collection of e-prints for Physics, Mathematics, Computer Science, Quantitative Biology, and Statistics. Accuracy was calculated as the sum of the differences between the predicted and true values for all articles. The winning entry used a k-nearest neighbors approach [57].

There have been a number of papers focused on the same prediction task besides this competition. Csárdi [58] predicted citation counts by studying the evolution of a citation network as documents were added to an existing network. The probability of a new paper citing a specific paper depended on its age and citation count. Recent papers with more citations were more likely to be cited. Csárdi replicated the KDD Cup prediction task although this method was not specifically designed for the prediction task. Probabilities were estimated from the initial document set, and the growth of the network was simulated. The number of citations was averaged over multiple simulations to determine the final prediction. Performance was slightly worse than the best performance from the KDD Cup. Castillo et al. [59] used linear regression and citation count after 6 months to predict citation count after 30 months. They incorporated author-related information (i.e., the number of previous citations, publications, and co-authors for an author) to improve predictions. The resulting model had a correlation coefficient of 0.81 between the true and predicted number of citations for 1500 articles from Citeseer, a database of computer science articles. The correlation coefficient was 0.57 without author information which demonstrated that author data improved prediction.

Lokker [60] recently presented a regression model for predicting citation counts two years after publication using information available within three weeks of publication. This study performed multiple regression on 17 article-specific features and 3 journal-

specific features. Nine article-specific predictors were statistically significant including the number of authors, the number of pages, the number of references, and whether the article was abstracted in an evidence-based medicine journal. Significant journal-specific features were the number of databases that indexed the journal and the proportion of articles that were abstracted. The training set contained 1274 articles published in 105 journals from January to June 2005. Lokker's model predicted 56% of the variation for a holdout test set ($R^2 = 0.56$). The sensitivity and specificity of the model were 83.3% and 71.5% for the top half of cited papers. The values were 66.1% and 82.2% for the top third. The area under the receiver operating characteristic curve (AUC) was 0.76 for a median threshold of 7 citations.

Other work focused on slightly different learning tasks. Feitelson [51] modeled the citation rate of authors with a multiplicative model. The link prediction task predicts new links in a fixed network of nodes without adding new documents [61, 62]. Popescul [61] used author names, the citation graph, publication venue, and word count to predict unobserved links for a fixed set of documents. Taskar [62] defined a probabilistic model over the link graph by applying a relational Markov network. He predicted the relation type of links within a website and friendship links between students in an online community. Liben-Nowell [63] studied social networks and predicted new collaborations between researchers. Similarity was measured with topological features such as the number of common neighbors or paths between nodes. Al-Hasan considered the same prediction task but considered non-topological features such as the number of previous publications by an author and paper keywords [64].

Despite the apparent difficulties in citation prediction, the goal of this work was to examine the feasibility of citation count prediction in the biomedical literature. Support vector machine models were trained with the full content terms of the MEDLINE abstract and MeSH keywords as well as bibliometric information about the authors, journals, and institutions. The topic and subject matter of the article were included because heavily cited topics may be predictive of future citations. Bibliometric features were included since social factors may affect citation rates. This approach differed from previous methods since only information available at publication time was used. In addition to the model-building effort, the models were analyzed to identify factors that correlate strongly and potentially determine the chances of an article being cited by many subsequent articles.

Methods

Predictive Features and Response Variable

Table 8: Features included in each model for citation count prediction

Feature	Complete model	Content model	Bibliometric model	Impact Factor model
Article title	x	x		
Article abstract	x	x		
MeSH terms	x	x		
Number of articles for first author	x		x	
Number of citations for first author	x		x	
Number of articles for last author	x		x	
Number of citations for last author	x		x	
Publication type	x		x	
Number of authors	x		x	
Number of institutions	x		x	
Journal impact factor	x		x	x
Quality of first author's institution	x		x	

Table 8 lists the input features used to construct a learning corpus for predictive modeling. The *number of articles or citations for first and last authors* was counted for 10 years prior to publication. *Publication type* indicates if a paper was identified as an article or review by the Institute of Scientific Information's (ISI) Web of Science (WOS) bibliometric database [44]. The Academic Ranking of World Universities (ARWU) [65] was used to measure the *quality for the first author's institution*. ARWU was compiled by the Shanghai Jiao Tong University and ranked the top 500 universities in the world. Ranking criteria included the number of Nobel Prize and Fields Medal recipients among its alumni and faculty, the number of "highly-cited researchers" according to ISI [66], number of articles published in Nature and Science, the number of articles indexed in the Science Citation Index and Social Science Citation Index, and the size of the institution.

Number of institutions refers to unique home institutions for all authors. All other variables are self-explanatory.

The response variable was defined by a set of citation thresholds to determine if an article was labeled positive or negative. For a given threshold, a positive label meant an article received at least that number of citations within 10 years of publication. Thresholds were chosen (before analysis) to be 20, 50, 100, and 500 citations. In the space of topics covered by the corpus (see next subsection), papers with at least 20, 50, 100, and 500 citations within 10 years can be interpreted to be at least mildly influential, relatively influential, influential, and extremely influential respectively.

Predictions were made for a binary response variable rather than a continuous one because error metrics for discrete values are easier to interpret than continuous ones. Continuous loss functions such as mean square error or percent variation explained are more difficult to interpret in terms of practical significance.

Corpus Construction

The corpus was built for model training and evaluation by specifying a set of topics, journals, and dates. Eight topics were chosen from internal medicine as defined by the MeSH vocabulary: Cardiology, Endocrinology, Gastroenterology, Hematology, Medical Oncology, Nephrology, Pulmonary Disease, and Rheumatology. An article was operationally considered relevant to a topic if its MEDLINE record contained one of the eight MeSH terms, a related topic from the "See Also" field of the MeSH record, or a term from a sub-tree of one of these terms [42]. For example, an article was Cardiology-

related if it contained the MeSH heading “Cardiology”, a related term like “Cardiovascular Diseases”, or a term from a sub-tree.

Articles were included from six journals: American Journal of Medicine, Annals of Internal Medicine, British Medical Journal, Journal of the American Medical Association, Lancet, and New England Journal of Medicine. The journals were selected to include popular journals with a broad range of impact factors. The corpus contained articles published between 1991 and 1994 to collect citation data for a 10 year period after publication of the most recent articles. The window length was chosen so that citation rates would have sufficient time to stabilize.

PubMed was queried for all desired articles, and additional information was downloaded from the bibliometric database, the ISI Web of Science (WOS) [44]. Documents were excluded if bibliometric data was unavailable, and the final corpus contained 3788 documents. The complete model consisted of 20005 total features, and information was downloaded in May 2007. Positive-to-negative class ratios for each threshold were as follows: 2705/1083 for threshold 20, 1858/1930 for threshold 50, 1136/2652 for threshold 100, and 100/3688 for threshold 500 citations. Figure 6 shows that citations followed a power law distribution where most papers received a small number of citations.

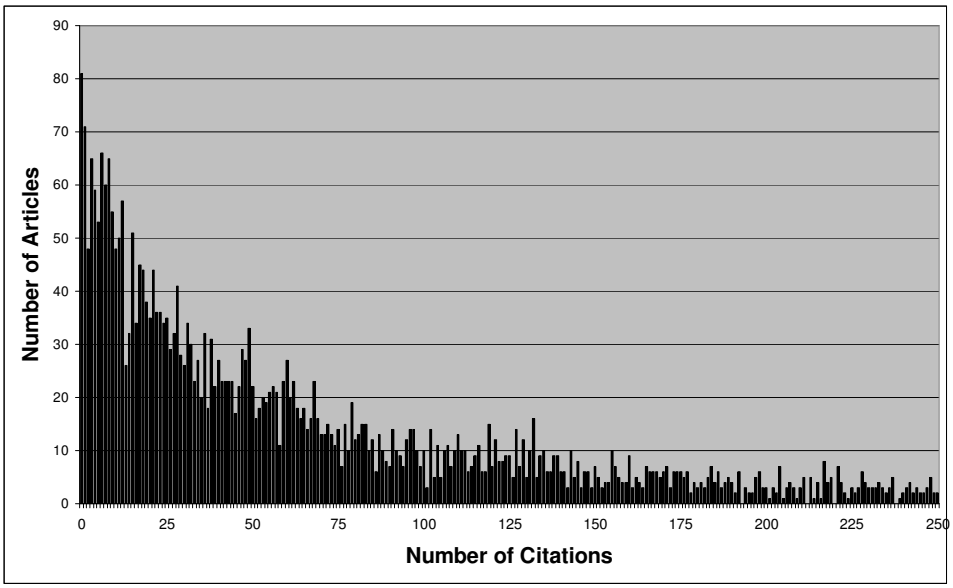

Figure 6: Distribution of citations for papers in the corpus (n = 3788 papers)

Document Representation

Articles were formatted for learning by text preprocessing and term weighting. The title, abstract, and MeSH terms were extracted from MEDLINE records. PubMed stop words (such as "the" or "a") [67] were removed from the title and abstract. Multiple forms of the same word were eliminated with the Porter stemming algorithm [68] to reduce the dimensionality of the input space.

Terms were weighted using log frequency with redundancy [69]. First, the number of times a term appeared in a document was transformed into a log frequency. Then it was multiplied by an importance weight (i.e. redundancy). Redundancy measured how uniformly distributed a term was throughout a corpus. A term appearing in all documents is not helpful for classification. A term appearing many times in one article while occurring once in each of the remaining articles is more discriminative [69].

The redundancy value for term k, r_k, is:

$$r_k = \log N + \sum_{i=1}^{N} \frac{f(w_k, d_i)}{f(w_k)} \log \frac{f(w_k, d_i)}{f(w_k)}$$

where N is the number of documents in the corpus, $f(w_k, d_i)$ is the number of occurrences of term k in document i, and $f(w_k)$ is the number of occurrences of term k in the corpus. The final step was L2-normalization to account for different text lengths. The vector of feature weights for a document i, x_i, is:

$$\mathrm{x}_i = \frac{\mathrm{l}_i * \mathrm{r}}{\left\| \mathrm{l}_i * \mathrm{r} \right\|_{L2}}$$

where l_i is a vector of the log frequencies for all terms in document i, r is a vector of redundancy values for all terms in the corpus, $\mathrm{l}_i * \mathrm{r}$ signifies component multiplication, and $\left\| \mathrm{l}_i * \mathrm{r} \right\|_{L2}$ is the L2-norm of the resultant vector. Each weight was a value between 0 and 1. In the end, the corpus was represented as a matrix where rows corresponded to documents and columns represented terms. Bibliometric features were scaled linearly between 0 and 1.

Learning Method

Support vector machine (SVM) models were used as the learning algorithm. They are a supervised learning method where a kernel function maps the input space to a higher-dimensional feature space, and a hyperplane is calculated to separate the classes of data [70]. The optimal hyperplane is the solution to a constrained quadratic optimization

problem. SVM models are usually sparse since the solution depends on the support vectors or points closest to the hyperplane [71]. Most features have zero weights, and the number of support vectors will be much smaller than the number of instances in most cases. This property makes SVMs suitable for representing text which typically involves high-dimensional data. Prior research has demonstrated that they perform well in categorizing text and identifying high-quality articles [13, 69]. In this application, input features were the weighted terms from MEDLINE records and the Web of Science.

Model Selection and Error Estimation

Models were selected with 5-fold nested cross validation. Parameters were optimized for cost and degree in the inner loop while the outer loop produced an unbiased estimate of model predictivity. The set of costs was [.1, .2, .4, .7, .9, 1, 5, 10, 20], and the set of degrees was [1, 2, 3, 4, 5, 8]. Performance was measured by area under the receiver operating characteristic curve (AUC). AUC was chosen instead of accuracy since AUC is not dependent on the ratio of positive and negative cases. Recall that an AUC of 0.5 describes a random classifier, AUC of ~.75 a mediocre classifier, AUC of ~0.85 a very good classifier, and AUC > 0.9 an excellent classifier (while an AUC of 1 denotes perfect classification).

Prospective validation was performed to analyze the models' ability to predict citation counts for future unseen articles. Articles from 1993 and 1994 were set aside for independent validation purposes, and articles from 1991 and 1992 were used to derive predictive models using the nested cross-validation procedure described.

Analysis of Influential Features

After fitting the complete models (i.e., with all features) and estimating their performance, the most influential features were identified using three types of analysis. First, reduced-feature models were trained for each threshold based only on the content, bibliometric data, or impact factor. Table 8 shows the features included in each model. Performance of these models revealed whether one type of feature was more important than the others.

A second feature-specific analysis was performed as follows: the total number of features was reduced by selecting the Markov Blanket of the response variable (i.e., number of citations received). The Markov Blanket is the smallest set of features conditioned on which all remaining features are independent of the response variable. It excludes irrelevant and redundant variables without compromising predictivity, and it provably results in maximum variable compression under broad distributional assumptions [72]. The specific algorithm used was semi-interleaved HITON-PC without symmetry correction which is an instance of the Generalized Local Learning class of algorithms [72]. It was verified that the reduced feature set predicted citation counts as well as the original model. After this variable selection and verification step, logistic regression estimated the magnitude of each feature's effect and its statistical significance on predicting citation counts *while controlling for all other features in the logistic regression model*. The raw SVM weights or Recursive Feature Elimination (RFE) weights in the polynomial SVM case cannot be used for the same purpose. SVMs do not control for the effect of all other variables on the weight of each feature in the SVM

model contrary to logistic regression. SVMs "spread" weights to otherwise conditionally independent features in order to implicitly model a smoother decision function.

The third method for identifying important features was SVM-based feature selection where features were ranked by linear SVM weights [73]. Features with the largest weights exert the greatest influence in defining the decision boundary. The majority of features have weights of zero, while the features with non-zero weights are support vectors. Cost was optimized for a linear SVM model, the model was re-trained, and each feature was ranked according to its linear SVM weight.

Implementation Details

Corpus construction and feature weighting were implemented with custom Python scripts. For text-based features, the scripts constructed PubMed queries, retrieved desired articles, downloaded MEDLINE records, and preprocessed text. For bibliometric features, the WOS database was queried with the title, author, and journal of each article. If a match was found, a user session was simulated by navigating through the website and extracting desired information about the document and authors.

The remainder of the code was written in MATLAB. LIBSVM was used to train SVM models, and it included a MATLAB interface [74]. Scripts were written to perform cross-validation and estimate performance. A custom MATLAB implementation for HITON was used as well as the logistic regression implementation of the MATLAB statistics toolbox.

Results

Overall Predictivity

Figure 7 shows the performance of four different types of models: the *complete model* with all features, models with *only content features*, models with *only bibliometric features*, and models with *only the impact factor*. The complete model accurately predicted whether a publication received a given number of citations for each citation threshold. AUC values range from 0.857 to 0.918 depending on threshold. The SVMs were able to learn useful models from the combination of content and bibliometric information.

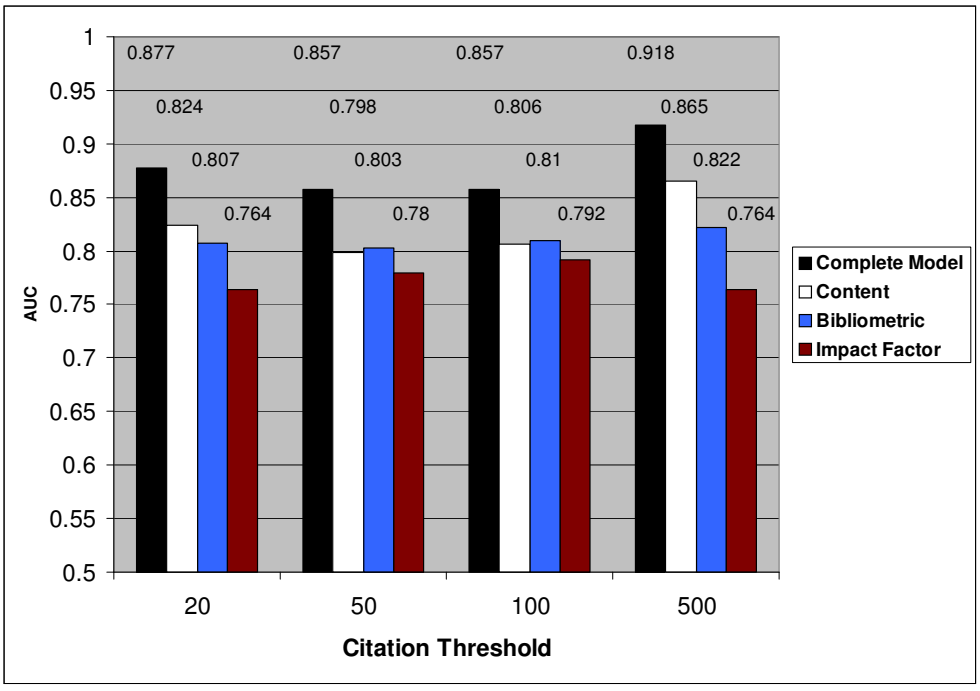

Figure 7: Performance for models based on all features, content, bibliometric features, and impact factor

Testing for Overfitting

Table 9: Cross-validation and prospective validation AUC results for citation count prediction

Citation Threshold	AUC cross-validated estimates (1991-1994)	Models built from 1991-2, tested on 1993-4
20	0.877	0.865
50	0.857	0.844
100	0.857	0.831
500	0.918	0.871

Table 9 shows the prospective validation results with the cross-validation estimates. Each row corresponds to a model for a given citation threshold. The second column shows estimated cross-validation performance in the full corpus (years 1991 to 1994). The third column shows performance of models built from years 1991-1992 when applied to documents from years 1993-1994. The models should generalize well since the cross-validation estimates are similar to the prospective validation results.

Another analysis was performed to further verify that the results were not overfitted. The method was borrowed from state-of-the-art analysis of high-throughput data [75]. Citation counts were randomly reshuffled, and all models were rebuilt on the reshuffled data exactly as was done for non-shuffled data. This procedure yielded AUC estimates of 0.5 since reshuffling eliminated the predictive association of the features to the outcome. This result verified that the original analysis was not overfitted.

Predictivity by Feature Type

After establishing that model performance was not due to overfitted analysis, the next analysis focused on estimating predictivity when learning with feature subsets. As

shown in Figure 7, the consistent trend in all thresholds was: AUC(complete model) ≥ AUC(content only features) ≥ AUC(bibliometric only features) ≥ AUC(impact factor only). Of the three reduced-feature models, no single model outperformed the other two for all thresholds. Impact factor had the lowest performance for all thresholds and was much lower than that of the complete model (differences in AUCs range from 0.065 to 0.154). The results also show that both content and bibliometric features had individually high predictivity. They both contributed to the accuracy of the complete model since AUC was maximized only when all types of predictive features were combined.

The Impact Factor model performed surprisingly well considering it is a poor predictor of citation count and does not correlate strongly with it [19]. In this corpus, the Pearson's correlation coefficient between Impact Factor and citation count was .429, and Spearman's rank correlation coefficient was .570. The predictive ability of Impact Factor was investigated by analyzing the distribution of citation counts for each journal shown in Figures 8-13. Citation counts followed a power law distribution for the entire corpus but did not retain this distribution in each journal. AJM and BMJ followed a power law distribution where most articles received a small number of citations. The distribution was more spread out for the other journals. NEJM had many articles that were highly cited. This behavior may partially explain why Impact Factor was a reasonably effective predictor by itself with this corpus.

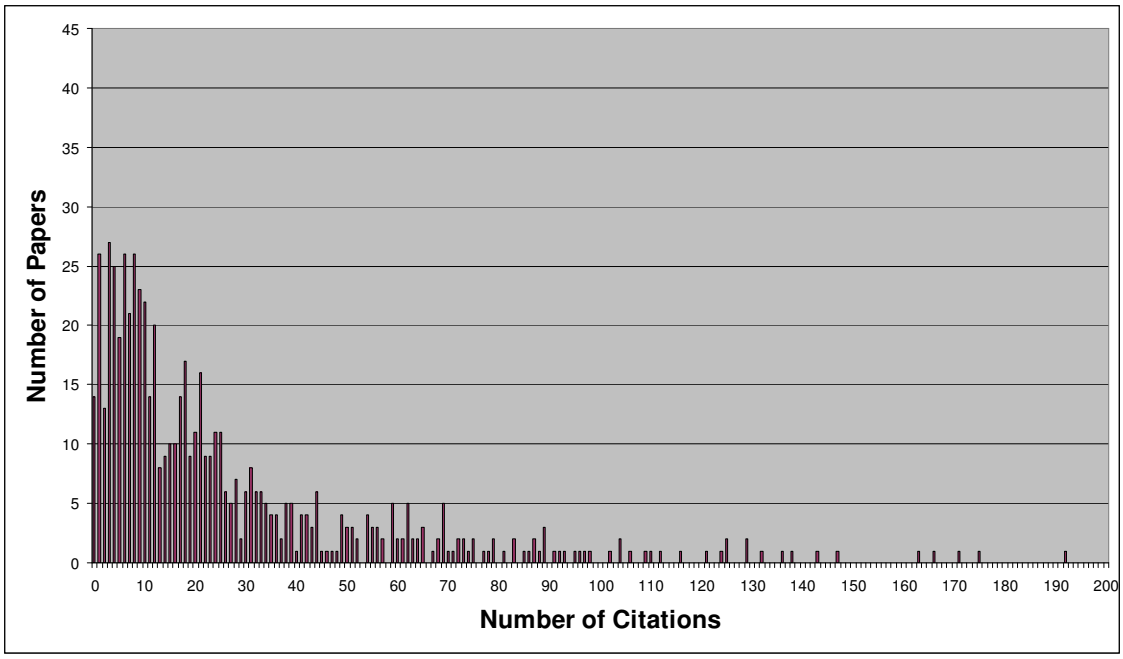

Figure 8: Distribution of citations over papers in American Journal of Medicine

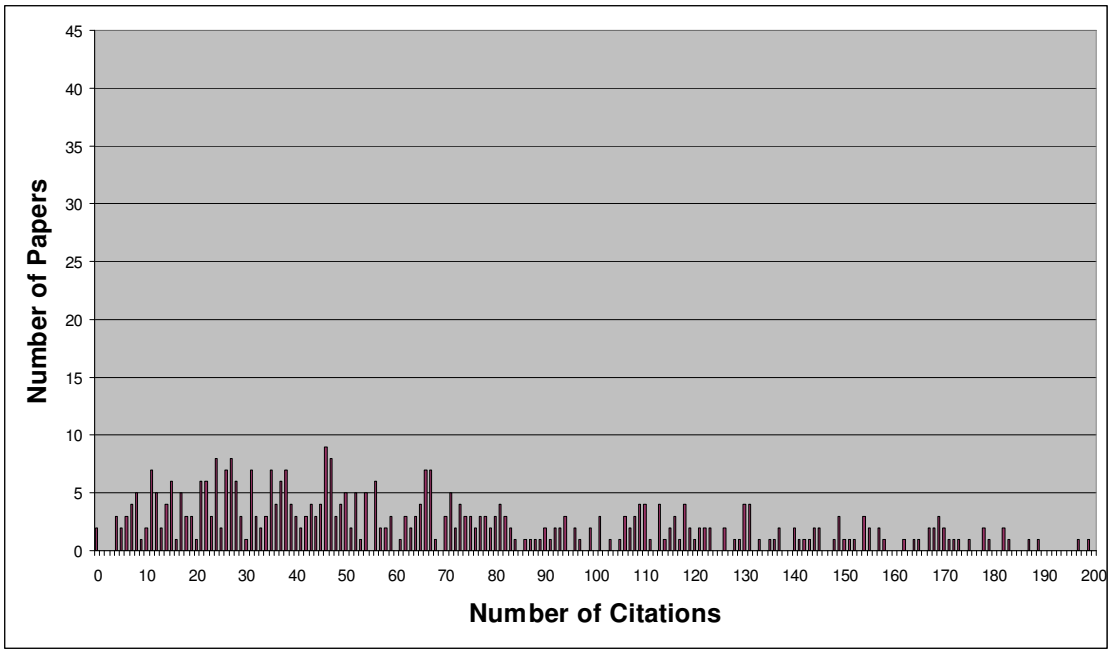

Figure 9: Distributions of citations over papers in Annals of Internal Medicine

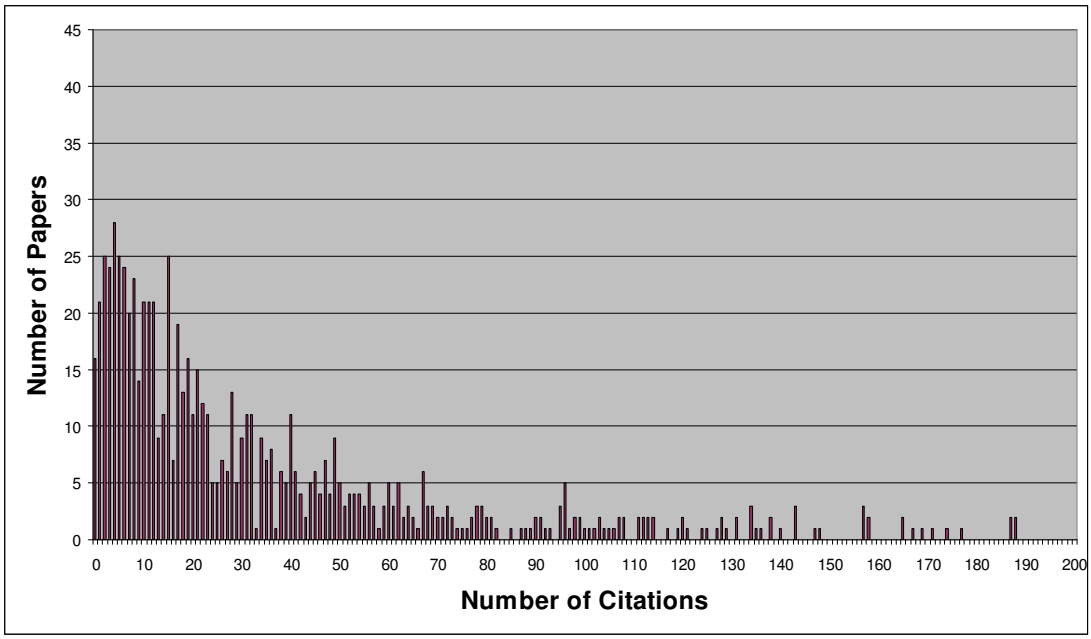

Figure 10: Distribution of citations over papers in the British Medical Journal

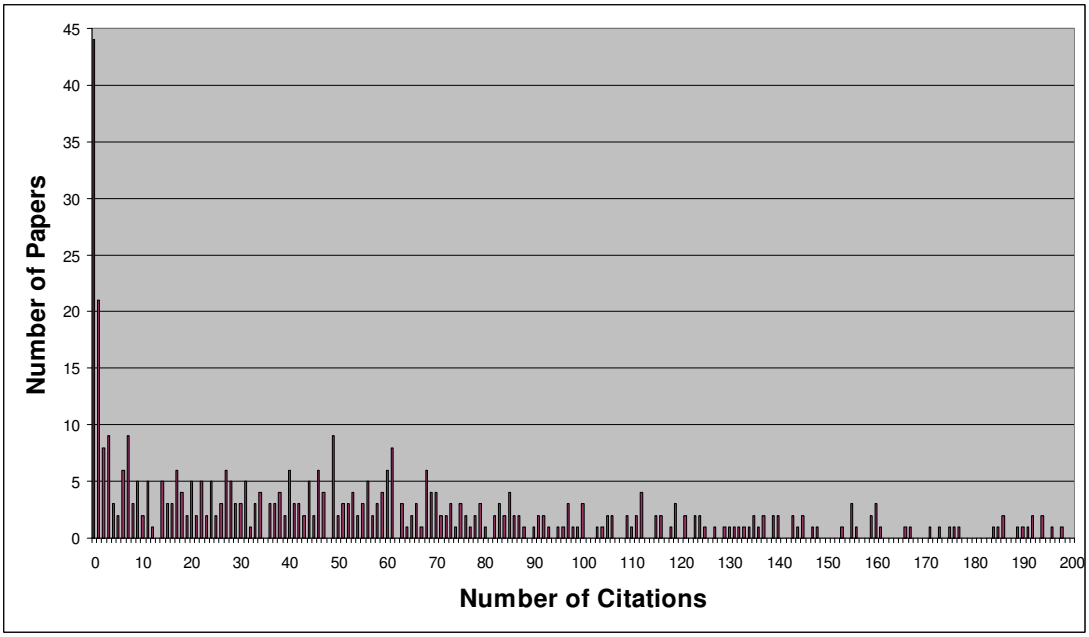

Figure 11: Distribution of citations over papers in JAMA

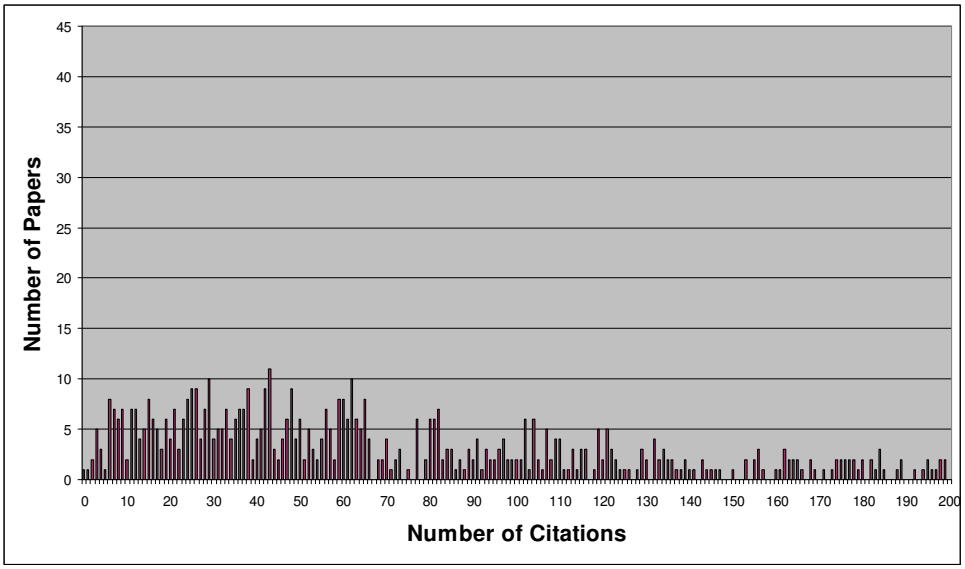

Figure 12: Distribution of citations over papers in Lancet

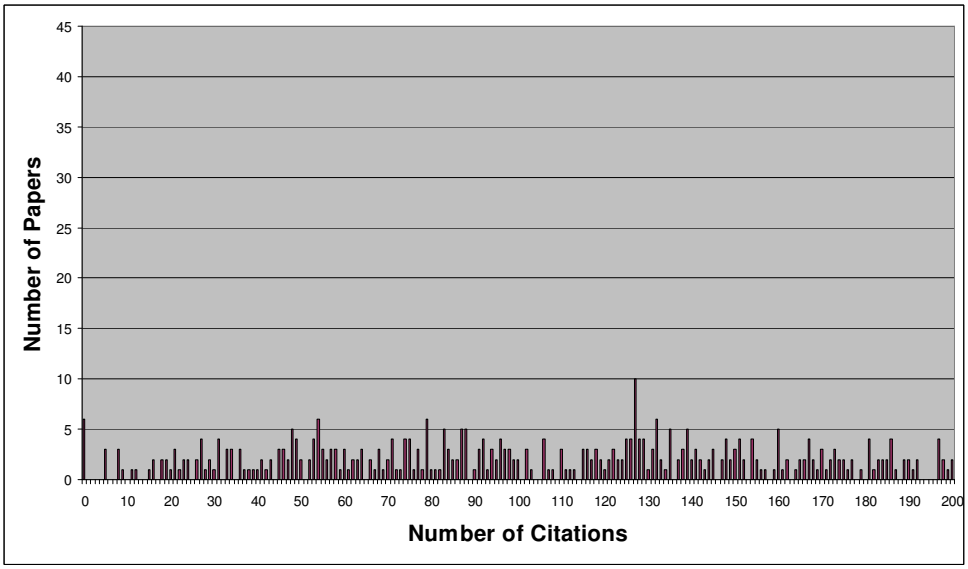

Figure 13: Distribution of citations over papers in New England Journal of Medicine

Analysis of Influential Features

As explained in the methods section, the Markov Blanket only includes non-redundant and relevant features, and logistic regression estimated feature importance and statistical significance of the selected features. The original set of 20,005 features was reduced to 169, 125, 132, and 138 features for thresholds 20, 50, 100, and 500 respectively. Performance did not degrade substantially when the HITON set of features was used for learning rather than the full set. Tables 10-12 show the top 25 ranked features according to absolute values of regression coefficients for citation thresholds 20, 50, and 100. Features with p-values greater than 0.05 were removed. There is no table for threshold 500 since all p-values were greater than 0.05. Features with the label "[MeSH]" were MeSH term headings in the MEDLINE records, features with "[Title]" were words from the title, and features with "[WOS]" were bibliometric features. Features without labels were terms from the abstract.

Recall that a positive unit change in a regression coefficient β for a feature corresponds to e^{β} increase in the odds of exceeding the citation count threshold for which the model is built. For example, "First Author Citations" had the largest coefficient of 5.753 for citation threshold 100. This value indicates that an article with the greatest number of first author citations was about 315 times ($e^{5.753} \approx 315$) more likely to receive 100 citations than an article with no first author citations. A one-unit change for interval-based features corresponds to a difference between the largest and smallest values since interval variables were scaled in the [0,1] range.

The feature-specific analysis points to several important conclusions: (a) certain "hot" topics were associated with high citation rates (e.g., smoking:mortality [MeSH]

was 68 times more likely to exceed 100 citations when controlling for other factors); (b) other topics or types of practice indicated smaller citation probability (e.g., splenectomi* and family practice were about 33 and 17 times less likely to receive 50 and 100 citations); (c) citation history of first and last author played a significant role in citation rates by increasing the chances of exceeding 100 and 50 citations by 315 and 23 times when comparing the best and worst citation histories; (d) For each threshold, different sets of content features were selected (and ranked differently in the top positions) which indicates that the importance of content changed for different levels of citation impact. On the other hand, bibliometric features and impact factor were predictive and always had large positive effects for all thresholds studied.

Table 10: Top 25 features sorted by absolute value of regression coefficient (threshold 20). A regression coefficient β for a feature corresponds to e^{β} increase in the odds of an article receiving more than 20 citations. "[WOS]" refers to bilbiometric features, "[MeSH]" refers to MeSH terms, "[Title]" refers to terms occurring in an article's title.

Feature	Regression Coefficient	P-value	Standard Error
Cardiac Tamponade [MeSH]	-4.939	0.000	1.282
splenomegali	-4.927	0.007	1.832
Journal Impact Factor [WOS]	4.040	0.000	0.252
supply & distribution [MeSH]	-3.966	0.002	1.257
ectopi	-3.585	0.007	1.324
Thrombocytopenia:immunology [MeSH]	-3.560	0.008	1.335
Internal Medicine [MeSH]	-3.537	0.001	1.023
Lung Neoplasms:etiology [MeSH]	-3.438	0.001	1.000
Cholelithiasis [MeSH]	-3.274	0.010	1.272
Kidney Failure, Chronic:metabolism [MeSH]	-3.108	0.004	1.087
Ventricular Fibrillation [MeSH]	-2.962	0.001	0.878
tomographi [Title]	-2.935	0.028	1.332
increment	2.892	0.040	1.411
gradual	-2.767	0.001	0.842
history [MeSH]	-2.688	0.003	0.891
Oxygen:blood [MeSH]	-2.655	0.024	1.180
tachycardia [Title]	-2.578	0.000	0.671
periton [Title]	-2.481	0.047	1.252
clinicopatholog [Title]	-2.424	0.011	0.952
Clinical Protocols [MeSH]	-2.096	0.017	0.878
sucralf	-2.029	0.002	0.645
european [Title]	-1.807	0.007	0.673
transmiss	1.792	0.022	0.783
present [Title]	-1.792	0.031	0.831
liver [Title]	-1.644	0.003	0.549

Table 11: Top 25 features sorted by absolute value of regression coefficient (threshold 50). A regression coefficient β for a feature corresponds to e^{β} increase in the odds of an article receiving more than 50 citations. "[WOS]" refers to bilbiometric features, "[MeSH]" refers to MeSH terms, "[Title]" refers to terms occurring in an article's title.

Feature	Regression Coefficient	P-value	Standard Error
splenectomi	-3.406	0.006	1.243
Journal Impact Factor [WOS]	3.342	0.000	0.164
Last Author Citations [WOS]	3.147	0.001	0.914
ciprofloxacin	-2.858	0.019	1.223
Anemia, Sickle Cell [MeSH]	-2.760	0.000	0.681
Rural Health [MeSH]	-2.668	0.015	1.097
brain	2.574	0.000	0.635
history [MeSH]	-2.442	0.046	1.227
Zidovudine:therapeutic use [MeSH]	2.424	0.030	1.114
Death, Sudden [MeSH]	-2.329	0.014	0.948
catecholamin	-2.210	0.026	0.996
uncompl	-2.167	0.014	0.884
hypoglycaem	-2.143	0.048	1.084
inappropri	-1.857	0.038	0.894
ambulatori [Title]	-1.777	0.020	0.765
took	1.708	0.003	0.574
Molecular Sequence Data [MeSH]	1.589	0.006	0.583
Atrial Fibrillation [MeSH]	1.567	0.010	0.612
pylori	1.522	0.007	0.566
output	-1.517	0.003	0.514
Article Type [WOS]	1.480	0.000	0.179
Pilot Projects [MeSH]	-1.355	0.033	0.637
chain	1.300	0.006	0.474
thrombosi	1.081	0.007	0.403
asthma	0.957	0.000	0.262

Table 12: Top 25 features sorted by absolute value of regression coefficient (threshold 100). A regression coefficient β for a feature corresponds to e^{β} increase in the odds of an article receiving more than 100 citations. "[WOS]" refers to bilbiometric features, "[MeSH]" refers to MeSH terms, "[Title]" refers to terms occurring in an article's title.

Feature	Regression Coefficient	P-value	Standard Error
First Author Citations [WOS]	5.753	0.000	1.469
Smoking:mortality [MeSH]	4.224	0.018	1.785
offset	3.347	0.007	1.232
Journal Impact Factor [WOS]	3.320	0.000	0.180
Last Author Citations [WOS]	3.023	0.001	0.872
Birth Weight [MeSH]	2.954	0.000	0.770
Pilot Projects [MeSH]	-2.912	0.013	1.173
Autoantibodies:blood [MeSH]	2.783	0.001	0.810
Family Practice [MeSH]	-2.746	0.016	1.140
gy	2.647	0.006	0.959
person [Title]	2.576	0.002	0.828
Mycobacterium tuberculosis [MeSH]	2.466	0.009	0.945
tran	2.458	0.041	1.203
Immunohistochemistry [MeSH]	2.375	0.011	0.931
Endothelium, Vascular [MeSH]	2.257	0.002	0.740
pylori	2.246	0.000	0.606
meta [Title]	1.947	0.002	0.637
quantifi	1.877	0.001	0.575
Kidney Diseases [MeSH]	-1.842	0.009	0.708
apolipoprotein	1.598	0.007	0.596
mutat [Title]	1.544	0.022	0.676
heparin	1.527	0.001	0.460
unselect	1.401	0.003	0.480
endogen	1.222	0.008	0.458
largest	1.183	0.043	0.586

A heatmap was created in Figure 14 and Figure 15 to visually display the relative importance of the features. The p-values were log transformed and negated to increase the spread of values. Features that were not present for a threshold were assigned a p-value of .05.

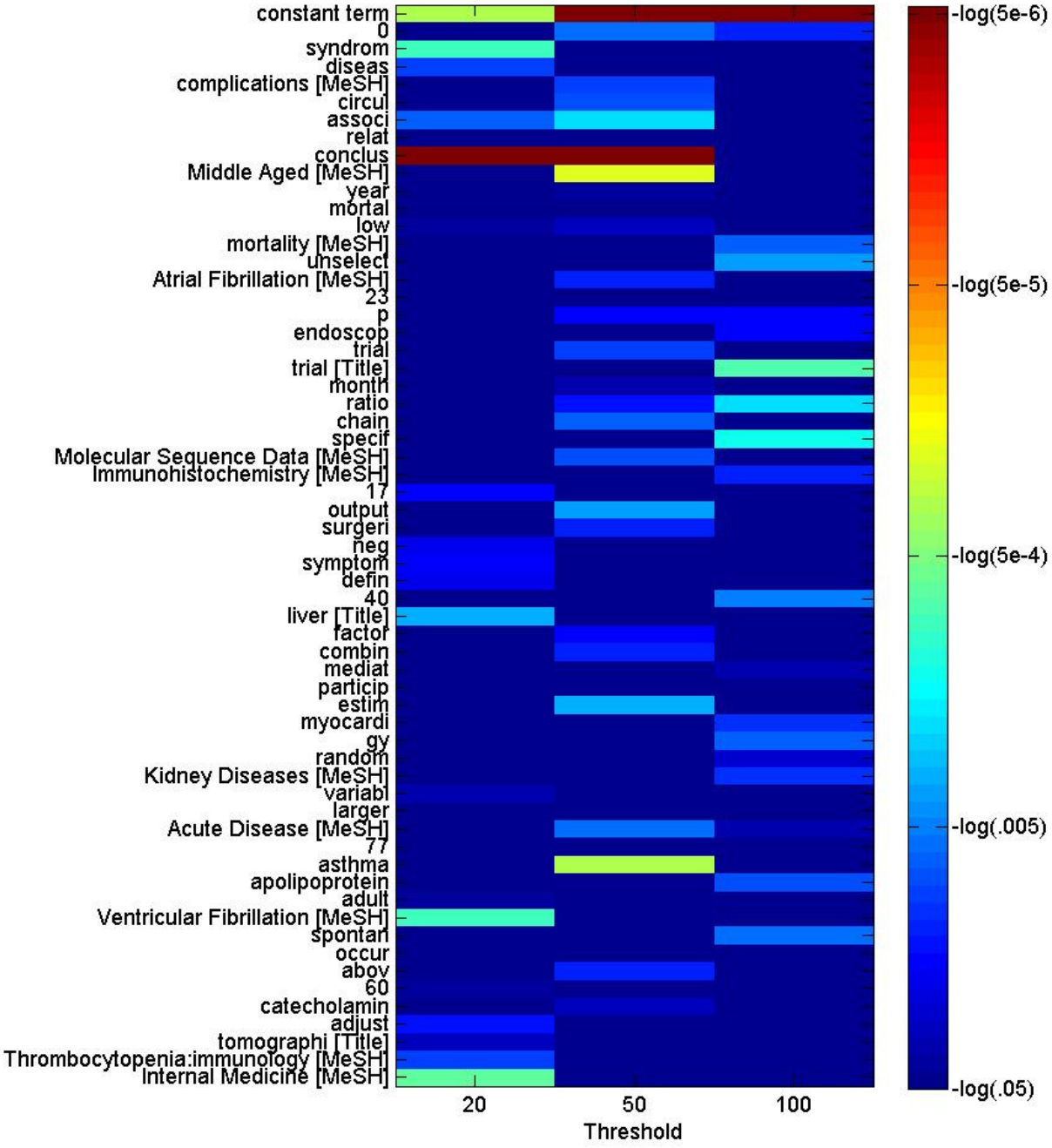

Figure 14: Heatmap of log transformed p-values (1 of 2)

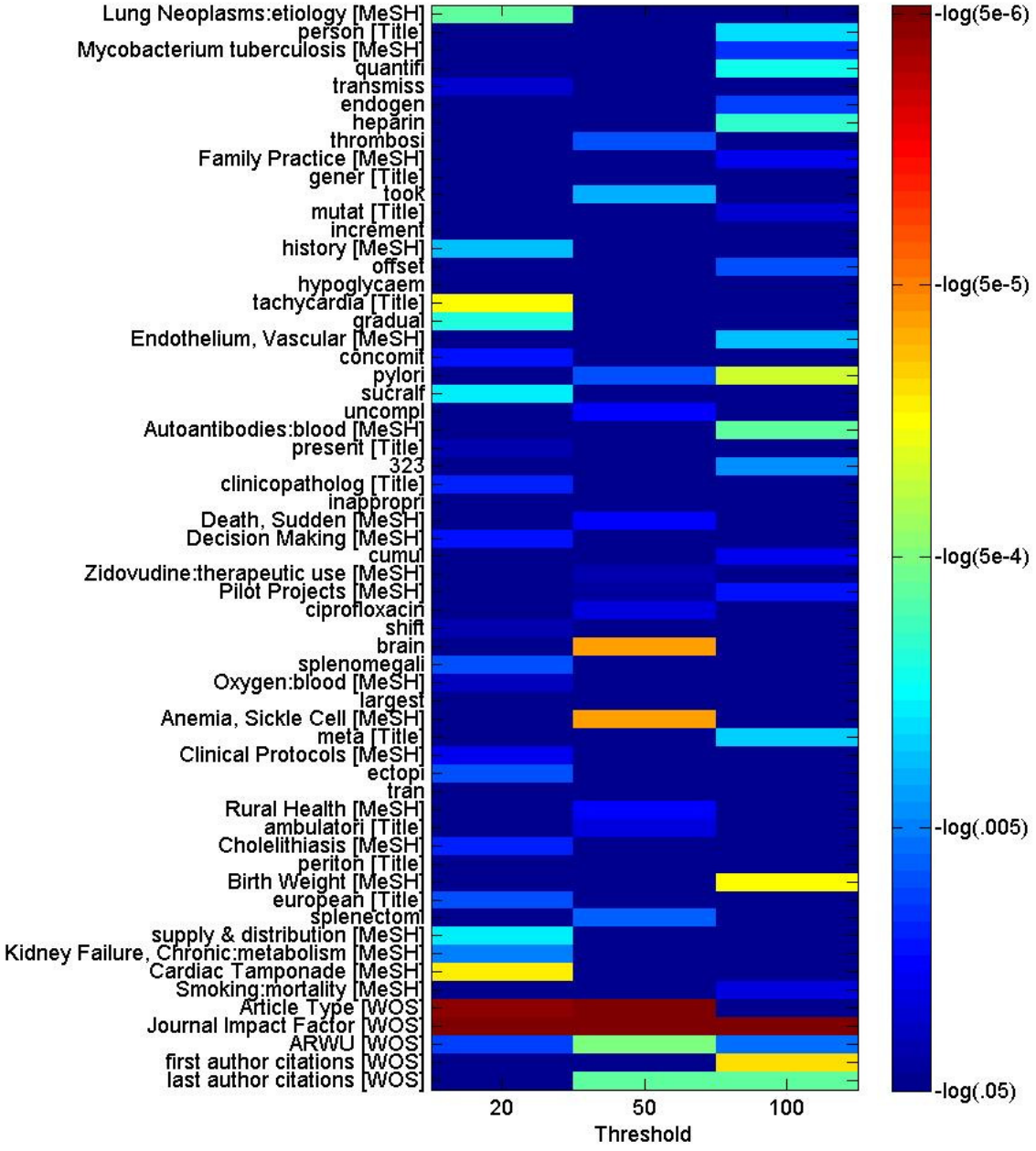

Figure 15: Heatmap of log transformed p-values (2 of 2)

Important features were also identified by ranking them with linear SVM weights. It was verified that performance did not degrade significantly when the degree parameter was not optimized. Tables 13-16 display the twenty-five features with the largest weights for each threshold. The SVM weights confirmed the importance of some of the features from the regression analysis. Content features such as “heparin” and “pylori” were influential in both sets of rankings. Also, bibliometric features such “Journal Impact Factor”, “Last Author Citations”, and “First Author Citations” were among the top features according to linear SVM weights.

Table 13: Top 25 features sorted by SVM weights (threshold 20). "[WOS]" refers to bilbiometric features, and "[MeSH]" refers to MeSH terms.

Feature	Weight
Journal Impact Factor [WOS]	1.898
object	1.734
main	1.681
outcom	1.505
design	1.487
associ	1.335
set	1.308
subject	1.230
symptom	1.181
measur	1.123
trends [MeSH]	1.115
receiv	1.100
infect	1.080
sucralf	1.079
occur	1.072
variabl	1.058
hospit	1.045
17	1.042
popul	1.040
antihypertens	1.027
conclus	1.019
increas	1.017
risk	1.016
patient	1.011
syndrom	1.003

Table 14: Top 25 features sorted by SVM weights (threshold 50). "[WOS]" refers to bilbiometric features, and "[MeSH]" refers to MeSH terms.

Feature	Weight
Journal Impact Factor [WOS]	1.867
associ	1.816
Last author citations [WOS]	1.672
p	1.277
object	1.256
outcom	1.206
stroke	1.189
ratio	1.157
ischaem	1.105
control	1.100
main	1.090
Number of institutions [WOS]	1.050
wave	1.039
adjust	1.007
tachycardia	1.006
trial [Title]	0.995
babi	0.984
cardiogen	0.968
1	0.960
mean	0.949
year	0.947
chain	0.936
statistics & numerical data [MeSH]	0.932
anti	0.924
Asthma:epidemiology [MeSH]	0.923

Table 15: Top 25 features sorted by SVM weights (threshold 100). "[WOS]" refers to bilbiometric features, and "[MeSH]" refers to MeSH terms.

Feature	Weight
percent	3.223
Last author citations [WOS]	1.927
First author citations [WOS]	1.672
drug	1.574
anti	1.545
Journal Impact Factor [WOS]	1.521
hcv	1.363
month	1.323
trial [Title]	1.318
heparin	1.315
particip	1.297
p	1.246
main	1.167
up	1.155
diseas	1.155
prostat	1.137
Number of authors [WOS]	1.135
low	1.122
1	1.119
randomis	1.102
odd	1.082
Evaluation Studies [MeSH]	1.076
specif	1.070
carri	1.057
allergen	1.044

Table 16: Top 25 features sorted by SVM weights (threshold 500). "[WOS]" refers to bilbiometric features, and "[MeSH]" refers to MeSH terms.

Feature	Weight
Number of authors [WOS]	0.133
Number of institutions [WOS]	0.102
estrogen	0.087
percent	0.072
pylori	0.064
c7e3	0.060
enalapril	0.058
Stomach Neoplasms:etiology [MeSH]	0.057
First author citations [WOS]	0.056
prostat	0.056
apc	0.055
Shock, Septic [MeSH]	0.053
gastric	0.052
grade	0.052
immedi	0.048
Pancreatic Neoplasms:therapy [MeSH]	0.048
Helicobacter Infections:complications [MeSH]	0.047
metastasi	0.046
concentr	0.046
reduc	0.044
intensifi	0.044
69	0.044
Tuberculosis [MeSH Main Heading]	0.043
placebo	0.043
Tuberculosis [MeSH]	0.042

Discussion

Limitations

The experimental corpus was restricted to internal medicine articles from 6 journals, and these articles were over ten years old. The limited coverage of topics and journals was chosen since this work was designed as a feasibility study. Also, the ten year window was chosen to allow sufficient time to pass so that citation rates would stabilize. Since the corpus only covered a small portion of the literature, it is unknown if the models will be useful for different time periods, journals, or topics. Citation prediction assumes that past citation behavior will remain unchanged in the future, and influential factors may no longer be relevant today. Over time, citation behavior may have changed significantly. For example, technological advancements have enabled the electronic distribution of the literature which may have changed how articles are cited. Also, the open access of journals could have affected citation behavior. Thus, the true usefulness of the models has yet to be determined.

Future Work

A logical continuation of this work would be studying the generalizability of the models. It is unknown how performance would change for different time periods, journals, and topics. Performance may improve with more recent publications. Shortening the timeframe for predicted impact would make this modification possible. Also, there are a number of possible refinements for improving the models. The number of publications or citations for an author could be weighted since some items may be a

better indicator of quality than others. Learning could improve by including the full text of articles into the content features along with the title and abstract [76]. In addition, alternative data sources could improve prediction. Citation databases such as Google Scholar and Scopus may be more accurate or comprehensive than the Web of Science [77]. Also, download counts and web access logs could be incorporated into the models since they may be useful in predicting impact [78].

Conclusion

The experiments showed that citation count can be accurately predicted for several distinct levels of citation performance with information strictly available at publication time. Recent developments in classifier technology have enabled the success of this method compared to previous approaches. These advances allow the use of all content terms in article titles, abstracts, and MeSH terms without suffering from the increased dimensionality. It is important to note that using content terms limits this method to journals indexed by PubMed.

This approach is very different from Lokker's method both in design and results [60]. Specifically, a longer time horizon was used for predictions, and a very large predictive feature space was used. Machine learning and feature selection algorithms identified predictive patterns while narrowing down the required features. Initial features included content and bibliometric information while Lokker's method used structural and systematic review criteria. The models produced in this work achieved predictivity that exceeded Lokker's model by about 0.10 to 0.16 AUC depending on the model. Notably, Lokker's model reported an AUC of 0.76 which should be no better (as evidenced in

these experiments with different feature sets) than a single relatively weak variable: the impact factor, which was not used in their models. Note that the results for the two studies cannot be conclusively compared because of the differences in chosen journals and time horizons. Because the two studies were independently conducted during roughly the same period,[1] the corpus and features set used by Lokker were not available for a direct comparison as part of this evaluation. This is clearly an area of interesting future research.

In conclusion, the results of the present work pave the way for practical models to predict future citations without requiring citations to build over time. Such models have the potential to render citation counts a more practical tool for evaluating long-term impact of recent work. Another advantage is providing an alternative to less accurate heuristics such as impact factor. Finally, analysis of the relative importance of various input variables for citation counts suggests that several factors may causatively influence or even bias citation practices.

[1] R.Brian Haynes, personal communication, November 2007

CHAPTER V

MACHINE LEARNING MODELS FOR AUTOMATIC CLASSIFICATION OF INSTRUMENTAL CITATIONS

Introduction

Evaluating the quality and impact of the scientific literature with citation count assumes that a citation is an indicator of quality. This is not necessarily true since a citation may serve many purposes unrelated to recognizing the value, rigor, or authority of the cited paper [17-19]. Cited papers may provide background information or acknowledge prior work that influenced the current work. Moreover, citations may serve non-scientific purposes due to social-psychological factors [16, 20, 21]. Thus, a citation is a subjective, indirect quality measure that does not have a single unambiguous use. For a more thorough discussion of the many motivating factors for a citation, see Chapter II.

Previous work has attempted to automatically classify citations according to the purpose of the citation [79-81]. Teufel automatically classified citation function based on cue phrases and a part-of-speech based recognizer [81]. Citations were assigned to one of twelve categories that reflected whether the citation described a weakness in the cited paper, compared or contrasted the work, praised or described an influential aspect of the work, or was neutral. The corpus contained conference articles in computational linguistics from the Computation and Language E-Print Archive (http://xxx.lanl.gov/cmp-lg), and the evaluation corpus contained 2829 citations from 116 articles. The corpus was manually labeled according to a classification scheme of 12

categories, and performance was evaluated by using the IBk algorithm as the learning method which is a k-nearest neighbor classifier. The results yielded Kappa and Macro-F values of .57, and percentage accuracy was .77. When the classifications were combined into the four general categories, Kappa was 0.59, Macro-F was 0.68, and percentage accuracy was 0.79.

Garzone and Mercer [79] proposed another method for automatically classifying citations. They believed that scientific writing utilizes certain phrases for persuasion that indicate the underlying rhetorical purpose of a citation and that citations can be classified with these phrases. Linguistic cues or phrases were manually identified from Physics and Biochemistry articles. For example, a citation in the results section containing the words "postulated", "reads", or "reported" was classified into a specific category. Their parser consisted of lexical rules based on cue words and grammar-like parsing rules to match sophisticated patterns. The classification scheme contained 35 categories with 195 lexical rules and 14 parsing rules.

Automatically classifying citations could improve citation indexers since the nature of the relationship between articles would be known. Researchers and users could determine if an article criticizes, praises, builds upon, or compares itself to a cited article [81]. Current indexers find articles citing a given article but would be more helpful if they could identify articles using similar techniques or ones presenting conflicting results [80]. Automatic classification could also make large databases of articles more manageable by identifying related articles and performing information extraction or text summarization [80].

Another potential benefit of classifying citations is improving citation metrics such as journal impact factor and article citation count. The performance of existing evaluation methods may improve if instrumental citations could be reliably distinguished from non-essential ones. For the purposes of this work, a citation was considered instrumental if either of the following rules were true: the hypothesis of the citing work was motivated by the cited work, or the citing work could not have been completed without the cited work. In other words, modified versions of citation count and journal impact factor could be better quality metrics if they only counted citations to papers that played a central role in the generation of the hypothesis or provided necessary foundational knowledge.

This portion of the thesis determined the feasibility of automatically differentiating between instrumental and non-instrumental citations using machine learning methods. The learning approach was similar to the one used for predicting citation count in Chapter IV. Support vector machine models were trained on content and bibliometric features. Content features included the citation text, title, abstract, and MeSH terms. Bibliometric features included the number of times a reference was cited in each section (i.e., introduction, methods, etc.) as well as the publication history of the first and last authors. Previous approaches used manually generated rules which can be labor intensive or subject to human bias. Machine learning models are automatically generated and not susceptible to these limitations.

This study was designed as a proof-of-concept with the potential to lead to later development of practical models if the method proved successful. The classification task was designed as a binary task for instrumental and non-instrumental citations since it

would be simpler than the one attempted by previous methods with multiple categories [79-81]. This choice was made since it was not known if machine learning models could effective classify citations or if the article content and bibliometric information provided useful information for classifying citations.

Methods

Definitions

This section provides specific definitions for terms used during subsequent discussion. An article that cites another work is called the *citing work*. The article that receives a citation is called the *cited work* or *reference*. A *citation* is the location in the text where a reference is cited which is typically denoted with a reference number in superscript or brackets. The *citation text* is the text surrounding the citation. Furthermore, a reference may be cited multiple times within the same article. Equivalently, a citing article may contain many citations to the same reference. The citation text for each citation is unique and consists of the text surrounding each citation.

For a specific example, consider the first citation in the introduction to this chapter: "This is not necessarily true since a citation may serve many purposes unrelated to recognizing the value, rigor, or authority of the cited paper [17-19]." This thesis is the citing work, and references [17-19] are the three cited works. The citation text is the sentence "This is not necessarily true…" The citation text can include any number of words before or after the citation.

A definition for an instrumental citation was required for labeling the corpus. For the purposes of the study, a citation was operationally defined as instrumental if either of the following rules was true for a citation:

I. The hypothesis of the citing work was motivated by the cited work

II. The citing work could not have been completed without the cited work

An example of a reference motivating the hypothesis of a work is shown in this excerpt [82]:

> Recently, it has been suggested that endothelium-dependent dilatation of resistance vessels in coronary and other vascular beds is impaired in hypertension and hypercholesterolemia[10,11,12,13]. Therefore, altered endothelium-dependent vasomotion of coronary resistance vessels may contribute to the cause of angina-like chest pain in patients with normal coronary arteries. The present study attempted to determine whether endothelium-dependent vasodilatation of coronary resistance vessels was impaired in patients with this syndrome.

In this case, the citing paper investigated whether endothelium-dependent vasodilatation of the coronary vasculature was impaired in patients with microvascular angina [82]. The citation text states that references 10-13 stimulated the hypothesis of the article and that the article builds on the cited work. Therefore, these citations were labeled instrumental.

For the second rule, there are many ways to interpret that a reference was necessary for completing a paper. A reference was instrumental if it provided foundational knowledge. A good example is reference 7 in an article investigating the connection between secondhand smoke and lung cancer [83]. The study exposed non-smokers to secondhand smoke and found metabolites of the tobacco-specific lung carcinogen NNK in their urine. A reference had shown that NNK induced tumors in rats: "NNK is a powerful pulmonary carcinogen, inducing predominantly adenocarcinomas in

the lungs of rats, mice, and hamsters regardless of the route of administration[5,6,7]" [83]. The relationship between NNK and lung cancer is necessary to prove the hypothesis of the citing work which makes the citation instrumental.

Other criteria for crucial references included if the citing work used the same experimental design or dataset as the references, addressed the weaknesses or limitations of prior work as part of its hypothesis, or used an experimental technique that was essential for completing the study. Also, the reference could have conducted related work involving other animals, diseases, or organ systems that led to findings applicable to the citing work.

An example of a non-instrumental citation was one related to a statistical method or computer software. These tools likely did not motivate the hypothesis, and the study probably could have been completed with alternative methods. Non-instrumental citations were also identifiable if the article explicitly made it clear that the cited work did not influence the hypothesis or the design of the study such as reference 28 in this citation: "We examined several potential mechanisms that might explain our results[27,28,29]" [84]. The citation indicated that the references were considered after the experiments were completed which meant they did not motivate the hypothesis and did not enable its testing.

Input Features and Response Variable

Table 17: Features included in models for automatically classifying citations

Feature	PubMed indexed reference	Non-PubMed indexed reference
Title of cited article	x	x
Abstract of cited article	x	
MeSH terms of cited article	x	
Citation text within citing article	x	x
Number of times cited in Introduction of citing article	x	x
Number of times cited in Methods of citing article	x	x
Number of times cited in Results of citing article	x	x
Number of times cited in Discussion of citing article	x	x
Citation count of cited article	x	
Number of articles for first author of cited article	x	
Number of citations for first author of cited article	x	
Number of articles for last author of cited article	x	
Number of citations for last author of cited article	x	
Number of authors for cited article	x	
Number of institutions for cited article	x	
Quality of first author's institution for cited article	x	

Table 17 lists the input features used to construct a learning corpus. The *citation text* included a window of 25 words before and after each citation for a total of 50 words. The *number of times a reference was cited in each section* was included since it could indicate the relative importance of a reference. For example, an essential reference may be cited more frequently in the discussion rather than the introduction or vice versa. The *citation count of the cited article* was calculated for 10 years after publication or until the citing article was published depending on whichever occurred first. For example, if the cited paper was published in 1981 while the citing paper was published in 1994, citations were counted for 1981-1991. If the cited paper was published in 1990 while the citing paper was published in 1994, citations were only counted from 1990 until 1993. This

adjustment ensured that only information available at publication time was used. The *number of articles or citations for first and last authors* was counted for 10 years prior to publication. The *number of institutions* refers to unique home institutions for all authors. The Academic Ranking of World Universities (ARWU) [65] was used as the measure of *quality for first author's institution*. All other variables are self-explanatory. PubMed and ISI did not index all references including books, reports, guidelines, and articles from some journals. In this case, references had input features of the article title, number of times cited in each section, and the citation text.

The response variable was determined by manual review. Each citation was labeled either instrumental or non-instrumental based on its relevance to the hypothesis of the citing work. The citation was labeled instrumental if the reference motivated the hypothesis or the citing work could not have been completed without the reference. More details were provided in the Definitions subsection of this Methods section.

Corpus Construction

The corpus was defined for a set of topics and dates. Eight topics were chosen to cover a wide range of topics from internal medicine as defined by the MeSH vocabulary: Cardiology, Endocrinology, Gastroenterology, Hematology, Medical Oncology, Nephrology, Pulmonary Disease, and Rheumatology. An article was relevant to a topic if its MEDLINE record contained one of the eight MeSH terms, a related topic from the "See Also" field of the MeSH record, or a term in a sub-tree of these terms [42]. For example, an article was Cardiology-related if its record contained the MeSH heading "Cardiology", a related term such as "Cardiovascular Diseases", or a sub-term of one of

these terms. The corpus consisted of all New England Journal of Medicine articles related to internal medicine that were published in 1993 and 1994. Articles from other journals were not included since the full text of articles was not accessible online for this time period.

The full text of the articles was downloaded from the New England Journal of Medicine website. Reviews and special articles without an obvious hypothesis were removed since it was not possible to identify instrumental citations according to the operational definition. Three references were randomly selected from each article, and all citations to these references were identified. Corresponding records were found in the Institute of Scientific Information (ISI) Web of Science (WOS) [44] if they were indexed, and all desired bibliometric information was downloaded. The final corpus contained 1310 citations from 272 articles. Each citation was manually reviewed and labeled as instrumental or non-instrumental according to the definition at the beginning of the Methods section of this chapter. The ratio of instrumental to non-instrumental citations was 949 to 361.

Document Representation and Learning Method

Articles were formatted with the same procedure used for predicting citation counts. Content terms were derived from the title, abstract, MeSH terms, and citation text. Stop words were removed, Porter stemming [68] was performed to remove multiple formats of the same word, and terms were weighted by log frequency with redundancy. For further details, refer to the Document Representation portion of the Methods section in Chapter IV.

Support vector machine (SVM) models were used as the learning method. As with the citation count prediction task, the models were trained with a combination of content and bibliometric features. Additional details were provided in the Learning Method portion of the Methods section in Chapter IV.

Model selection and error estimation

Models were selected with 5-fold nested cross validation. Parameters were optimized for cost and degree in the inner loop while the outer loop produced an unbiased estimate of model predictivity. The set of costs was [.1, .2, .4, .7, .9, 1, 5, 10, 20], and the set of degrees was [1, 2, 3, 4, 5, 8]. Performance was measured by area under the receiver operating characteristic curve (AUC).

Experiments were repeated with 3 variations. First, the corpus was separated by publication year (i.e., articles from 1993 and 1994) to see if performance was significantly different between the two years. Second, a hold out data set was excluded before training. Cross-validation and model training were performed on the training examples, and performance was evaluated on the hold out set. The hold out set was randomly selected as 30% of the citations, and results were averaged over 5 runs. Prospective validation was also performed where the models were trained on the 1993 articles and tested on the 1994 articles. The results for the hold out sets and prospective validation indicated whether the models are able to classify citations in unseen articles. If these results were similar to the cross-validation results, the models should be able to handle unseen cases.

The third experimental variation was randomly selecting one citation per reference and excluding the remaining citations from the analysis. This decision ensured that the data was independently and identically distributed. In the original experiments, citations to the same reference could occur in the training set as well as the testing set. This could be problematic since citations to the same reference are not independent. A citation is more likely to be instrumental if another citation to the same reference is instrumental. Furthermore, citations from the same reference would never occur in both the training set and unseen articles. This restriction resulted in a corpus of 816 citations.

Analysis of Influential Features

Table 18: List of features included in the content and bibliometric models

Feature	Content Model	Bibliometric Model
Article title	x	
Article abstract	x	
MeSH terms	x	
Citation text	x	
Number of times cited in Introduction		x
Number of times cited in Methods		x
Number of times cited in Results		x
Number of times cited in Discussion		x
Citation count of reference		x
Number of articles for first author		x
Number of citations for first author		x
Number of articles for last author		x
Number of citations for last author		x
Number of authors		x
Number of institutions		x
Quality of first author's institution		x

After estimating the model's performance in classifying instrumental citations, influential features were identified using two methods. First, reduced-feature models were trained only on the content or bibliometric data. Table 18 shows the features included in each model. Performance of these models would reveal if one type of feature was more important than the other. The second type of analysis involved Markov Blanket induction and logistic regression. The Markov Blanket excludes irrelevant and redundant variables to produce a reduced set of features. Logistic regression analysis estimated for each feature the magnitude of its effect and statistical significance while controlling for all other features in the logistic regression model. For further details, refer to the Analysis of Important Features portion of the Methods section of Chapter IV.

Implementation Details

Corpus construction and feature weighting were implemented in custom Python scripts. For text-based features, the scripts constructed PubMed queries, retrieved desired articles, downloaded MEDLINE records, and preprocessed text. For bibliometric features, the WOS database was queried with the title, author, and journal of each article. If a match was found, a user session was simulated by navigating through the website and extracting desired information about the document and authors.

The remainder of the code was written in MATLAB. LIBSVM was used to train SVM models, and it included a MATLAB interface [74]. Scripts were written to perform cross-validation and estimate performance. A custom MATLAB implementation for HITON was used as well as the logistic regression implementation of the MATLAB statistics toolbox.

Results

Classification Performance

Table 19: Cross-validation AUC results for the classification of citations experiments

Corpus	Cross-validation AUC	Hold Out Test Set AUC
Full Corpus	0.858	0.846
1993 articles	0.867	0.842
1994 articles	0.814	0.812
Train 1993, Test 1994	N/A	0.776

The cross-validation results in Table 19 show that it is possible to accurately classify instrumental citations. The model trained on the full corpus had an AUC of 0.858. Comparable performance was shown when the corpus was split up by year. AUC values were 0.867 and 0.814 for the 1993 and 1994 articles. Additional experiments were performed which excluded test cases before learning. Performance decreased slightly when a hold out test set was used. Cross-validation results decreased from 0.858 to 0.846 for the full corpus, from 0.867 to 0.842 for 1993 articles, and from 0.814 to 0.812 for 1994 articles.

The slight overfitting probably resulted from excluding test set information during feature weighting and scaling. Cross-validation weighted and scaled features with all corpus items without excluding the test set. For the text features, feature weighting calculated term distributions for redundancy values. Cross-validation included the test set in these computations while hold-out experiments did not. For the bibliometric data,

cross-validation scaled features over the range of values for all articles, while hold out experiments only considered training cases.

Another observation is that the models appear to be time dependent. Performance decreased between the 1993 and 1994 articles. Also, there was a larger performance decrease when training on 1993 articles and testing on 1994 articles. It is unclear if a larger training corpus would make the models more robust over time or if the models need to be built on a yearly basis.

Table 20: Results for classification of citations after restricting corpus to one citation per reference

Corpus	All Citations (AUC)	1 Citation per Reference (AUC)
Full Corpus	0.858	0.815
1993 articles	0.867	0.858
1994 articles	0.814	0.770

In the previous experiments, it was possible for citations to the same reference to occur both in the training and testing sets. Experiments were repeated after limiting the corpus to one citation per reference. Table 20 shows that learning with one citation per reference reduced classification performance. Cross-validation results decreased from 0.858 to 0.815 for the full corpus, from 0.867 to 0.858 for 1993 articles, and from 0.814 to 0.77 for 1994 articles. This finding is not surprising since citations to the same reference are not independent, and classification is probably easier when citations to the same reference occur in both the training and testing sets.

Analysis of Influential Features

Learning was performed on feature subsets to investigate whether content or bibliometric features were more important for classification. AUC performance was 0.858 for the complete model, 0.827 for the content model, and 0.771 for the bibliometric model. The content model slightly outperformed the bibliometric model, but they both performed relatively well in isolation. It appeared that both types contributed to the accuracy of the complete model.

Another method for studying influential features involved Markov Blanket induction and Logistic Regression. Markov Blanket induction selected only non-redundant and relevant features, and Logistic Regression estimated feature importance and statistical significance of the selected features. Cross-validation with the full corpus yielded 12912 features which were reduced to 67 features. Performance did not degrade substantially when learning with the HITON set of features. Table 21 ranks the features by absolute values of regression coefficients. Features with p-values greater than 0.05 were removed. Features with the label "[MeSH]" were MeSH term headings in the MEDLINE records, and features with "[WOS]" were bibliometric features. Features without labels were terms from the abstract or citation text.

Table 21: Top features sorted by absolute value of regression coefficients. A regression coefficient β for a feature corresponds to e^{β} increase in the odds of a citation being considered instrumental. "[WOS]" refers to bilbiometric features, and "[MeSH]" refers to MeSH terms.

Feature	Regression Coefficient	P-value	Standard Error
Number of times cited in introduction[WOS]	5.650	0.0000	0.704
24	-4.634	0.0005	1.338
von	-3.418	0.0088	1.305
mammographi	-2.902	0.0204	1.252
Cytarabine[MeSH]	-2.699	0.0008	0.804
Arrhythmias, Cardiac [MeSH]	-2.428	0.0127	0.974
complex	-2.380	0.0007	0.705
eject	-2.195	0.0027	0.732
visual	-1.966	0.0023	0.645
underestim	-1.891	0.0058	0.686
classification[MeSH]	-1.813	0.0011	0.556
vari	-1.556	0.0143	0.635
adjust	-1.278	0.0380	0.616
comparison	-1.264	0.0058	0.458
genetics[MeSH]	0.991	0.0097	0.383
mean	-0.987	0.0405	0.482
3	-0.921	0.0141	0.375
model	-0.905	0.0479	0.457
test	-0.840	0.0072	0.313
two	-0.695	0.0084	0.264
Female[MeSH]	0.511	0.0070	0.189
studi	0.406	0.0382	0.196

A positive unit change in a regression coefficient β for a feature corresponds to e^{β} increase in the odds of being an instrumental citation. For example, "Number of times cited in introduction" had the largest coefficient of 5.650. This value indicates that a reference with the most citations in the introduction was about 284 times ($e^{5.65} \approx 284$) more likely to be instrumental than one with no citations in the introduction. A one-unit change for interval-based features corresponds to a difference between the largest and smallest values since interval variables were scaled in the [0,1] range. The majority of

the features were negatively correlated with an instrumental citation. Features with positive associations included "genetics [MeSH]," "Female [MeSH]," and "studi," but they had smaller effects than the number of times a reference was cited in the introduction.

Discussion

Limitations

The experimental design was limited by the fact that the corpus contained only articles from one journal. Results and conclusions may not hold true for other journals. This restriction was due to the unavailability of full text articles for many journals during the studied time period. Another limitation was that the corpus was labeled by a single individual. Multiple subjects were not used since manually labeling the corpus required a significant amount of time and effort. However, the important result of the study is that the SVM models were able to accurately classify citations according to the provided gold standard. In this case, the gold standard was the individual rater's notion of an instrumental citation. In the future, it would be interesting to determine the method's ability to model another gold standard.

Future Work

Important future work would be to thoroughly evaluate the generalizability of the learning method by increasing the scope of the corpus. A larger corpus with articles from different journals, longer time periods, and more topics would be useful in evaluating the

ability of the models to classify instrumental citations. Other possible work could be studying whether other categorizations can be learned besides instrumental vs. non-instrumental citations. For example, it could be useful to identify all negative citations or citations contrasting the cited work. If the models can handle other categories, they may be able to classify citations into multiple categories instead of two.

The motivation for classifying citations was to improve citation indexers and citation metrics. The models should be integrated into citation indexers to determine if they can automatically identify articles that are related, use similar techniques, or contrast the citing work. The ability of the models for other related tasks such as information extraction and text summarization should also be investigated. Also, modified versions of journal impact factor and citation count should be computed by using the models to ignore non-instrumental citations. These modified versions should be compared to other accepted impact measures to see how well they correlate. Also, modified metrics could be computed using the classification schemes by Teufel and Mercer [80, 81], and their performances could be compared to modified metrics based on the SVM models.

Conclusion

The learning method presented in this work was significantly different from previous methods for automatic classification of citations. Teufel and Mercer [80, 81] devised methods based on human review of articles to generate rules, phrases, and cues for identifying citations. These methods are labor-intensive and subject to human bias or error. The models presented in this work are automatically generated and avoid these limitations.

This work was successful in demonstrating the feasibility of machine learning methods for automatically classifying instrumental citations. SVM models analyzed the textual content of articles along with bibliometric data to classify instrumental citations in a manually labeled corpus. Efforts were made to study the generalizability of these models and their ability to classify unseen instances. The results were encouraging, but further work is necessary to see if practical tools can be developed to improve journal impact factor and citation count in real-world applications.

CHAPTER VI

DISCUSSION

Summary of Results

The purpose of this dissertation was to improve the usability and performance of existing information retrieval techniques in biomedicine with machine learning methods. The first focus was analyzing evaluation methods for journals, articles, and websites and measuring the variability of their performance for specific topics. Query-independent methods such as journal impact factor, clinical query filters, and PageRank were relatively unstable for different topics. Topic-specific impact factor and SVM-based models were less sensitive to topic. It is important for users to be aware of this issue since topic-sensitive methods could provide misleading conclusions and lead to a flawed evaluation of the literature. Methods that consider the topic of the query or are insensitive to topic should be used whenever possible.

The second focus was examining the feasibility of predicting citation count with SVM models which could evaluate an article at the time of publication. Models were trained on the article content as well as bibliometric data. These models were able to accurately predict whether an article would surpass a given citation count for a range of thresholds. Experiments with reduced feature sets showed that both the content and bibliometric features contributed to the accuracy of the models. Unique content features were influential for different citation thresholds, and important bibliometric features included journal impact factor and the number of citations received by the first and last

authors. Prospective validation was performed where models were evaluated on examples that were excluded during training. Results were comparable to cross-validation results which suggest that the models can predict citation counts for unseen articles.

The third focus was investigating the ability of SVM models to automatically classify instrumental citations. This could increase the functionality of citation indexers and improve citation metrics such as journal impact factor and citation count by excluding unimportant citations. Models were trained on content and bibliometric features, and citation text was incorporated into the content features. A manually labeled corpus was used for evaluation. Citations were considered instrumental if the cited work motivated the hypothesis of the citing work, or the citing work could not have been completed without the cited work. Additional experiments were conducted by excluding test cases prior to model induction as well as restricting the corpus to one citation per reference. In all cases, SVM models were capable of classifying instrumental citations in the manually labeled corpus.

Limitations

Although this work provided encouraging results, more generalizability studies are needed. Efforts were made to evaluate the models on unseen cases, but the experimental corpus was restricted to a small number of journals, topics, and years. The citation behavior within this subset of the literature may differ from the literature as a whole. Results and conclusions from this work may not necessarily apply to other articles, and additional experiments should be conducted with articles from a larger

collection of journals for a wider range of dates and topics. Also, training sets should include more recent articles since citation behavior may have changed over time. If the results of these new experiments are consistent with the findings presented here, this would provide strong evidence that the models are truly generalizable.

Through the course of this work, it became apparent that the data sources dictated what types of experiments could be completed. For example, the corpus for classification of instrumental citations only included one journal because full text articles were not available for other journals during the time period studied. Also, the set of candidate features was limited by the available features in the Web of Science. Bibliometric research would benefit greatly if there were a citation database or data repository designed for research purposes that could handle large queries. It required a significant amount of time and effort to write the code to collect the data from the Web of Science and download the information.

Future Work and Open Questions

The models for predicting citation count and classifying instrumental citations were a first attempt at demonstrating the feasibility of the learning approach with the given input features. SVM models were previously used to identify high-quality articles. However, it was unknown if they could predict citation count or automatically classify citations. Furthermore, it was not known if the article content and bibliometric information were suitable input features for the prediction and classification tasks. Since this work has shown that the learning method and features are suitable for the task, the performance of the models may improve with further refinements such as incorporating

additional features. For example, the full-text of an article could be incorporated into the content terms of an article when predicting citation count. Download counts and web access logs have been shown to correlate with impact, and they could potentially provide useful information for the models.

Along with the prior work in identifying high-quality articles, the success of the SVM models here provides strong support for the suitability of SVM models for text categorization tasks in general. The models are able to handle high-dimensional data and combine multiple types of data (i.e., content and bibliometric data). Furthermore, the learning method could be useful for other learning tasks related to citation analysis.

In addition to the generalizability of the models and improvement of their performance, there are a number of open questions to investigate. The models should be compared directly to the methods of Lokker [60], Teufel [81], and Mercer [80]. Bornmann's review of citation analysis noted that many studies varied widely in design, presented unreliable results that could not be replicated, and suffered methodological weaknesses [16]. In order to compare the machine learning methods to alternative methods, their performances should be compared directly to each other on the same corpus for an identical learning task with the same evaluation metric.

There are a number of considerations that need to be solved to develop practical tools based on these models for regular use. The experiments were conducted on a static subset of the literature, and applying the models in real-world situations will present new complications. The durability of the models is unknown since important features may change over time. It will be necessary to figure out how to update the models with new cases as more articles are published. For example, how often should the models be

updated? Should the training set be limited to a number of recent years or include as many past articles as possible? If the training set is limited, how many years should be included? There are many considerations that need to be resolved to optimize the models with respect to performance and efficiency.

Conclusion

The main goal of this work was to improve the usability and performance of citation metrics for information retrieval within the biomedical literature by applying machine learning methods. This work raised awareness of the topic-sensitivity of several evaluation methods. Furthermore, it demonstrated the feasibility of SVM learning with content and bibliometric features for predicting citation count and classifying instrumental citations. The models appeared to generalize for some unseen cases, but additional experiments need to be performed on more journals, topics, time periods before general conclusions can be made. The results of this work indicate that it may be possible to develop practical applications and tools for use by researchers, clinicians, and consumers.

REFERENCES

1. Guyatt G, Cook D, Haynes R. Evidence based medicine has come a long way. BMJ. 2004. 329: 990-991.

2. Sackett D, Rosenberg W, Gray J, et al. Evidence based medicine: what it is and what it isn't. BMJ. 1996. 312: 71-72.

3. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet. 2006. 7(2): 119-129.

4. Loging W, Harland L, Williams-Jones B. High-throughput electronic biology: mining information for drug discovery. Nat Rev Drug Discov. 2007. 6(3): 220-230.

5. D'Alessandro DM, Kreiter CD, Peterson MW. An Evaluation of Information-Seeking Behaviors of General Pediatricians. Pediatrics. 2004. 113(1): 64-69.

6. Demner-Fushman D, Hauser SE, Humphrey SM, et al. *MEDLINE as a source of just-in-time answers to clinical questions*. in *Proceedings of the AMIA Annual Symposium*. 2006.

7. Ely JW, Osheroff JA, Ebell MH, et al. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. BMJ. 2002. 324(7339): 710-.

8. Magrabi F, Coiera EW, Westbrook JI, et al. General practitioners' use of online evidence during consultations. International Journal of Medical Informatics. 2005. 74(1): 1-12.

9. McKibbon KA, Fridsma DB. Effectiveness of Clinician-selected Electronic Information Resources for Answering Primary Care Physicians' Information Needs. J Am Med Inform Assoc. 2006. 13(6): 653-659.

10. Westbrook JI, Coiera EW, Gosling AS. Do Online Information Retrieval Systems Help Experienced Clinicians Answer Clinical Questions? J Am Med Inform Assoc. 2005. 12(3): 315-321.

11. Westbrook JI, Gosling AS, Coiera E. Do Clinicians Use Online Evidence to Support Patient Care? A Study of 55,000 Clinicians. J Am Med Inform Assoc. 2004. 11(2): 113-120.

12. Garfield E. The history and meaning of the journal impact factor. JAMA. 2006. 295(1): 90-93.

13. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, et al. Text categorization models for high-quality article retrieval in internal medicine. JAMIA. 2005. 12(2): 207-216.

14. Haynes R, Wilczynski N, McKibbon K, et al. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc. 1994. 1: 447-458.

15. Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Proceedings of the 7th International World Wide Web Conference. 1998.

16. Bornmann L, Daniel H. What do citation counts measure? A review of studies on citing behavior. Journal of Documentation. 2007.

17. MacRoberts M, MacRoberts B. Problems of citation analysis. Scientometrics. 1996. 36(3): 435-444.

18. Phelan T. A compendium of issues for citation analysis. Scientometrics. 1999. 45(1): 117-136.

19. Seglen P. Citation rates and journal impact factors are not suitable for evaluation of research. Acta Orthop Scand. 1998. 69(3): 224-9.

20. Cronin B. Metatheorizing Citation. Scientometrics. 1998. 43(1): 45-55.

21. Nicolaisen J. The Social Act of Citing: Towards New Horizons in Citation Theory. Proceedings of the 66th ASIST Annual Meeting. 2003. 12-20.

22. Garfield E. Can citation indexing be automated? Essays of an Information Scientist. 1962. 1: 84-90.

23. Moravcsik MJ, Murugesan P. Some Results on the Function and Quality of Citations. Social Studies of Science. 1975. 5(1): 86-92.

24. Chubin DE, Moitra SD. Content Analysis of References: Adjunct or Alternative to Citation Counting? Social Studies of Science. 1975. 5(4): 423-441.

25. Brooks TA. Private acts and public objects - an investigation of citer motivations. Journal of the American Society for Information Sciences. 1985. 36: 223-229.

26. Cano V. Citation behavior: Classification, utility, and location. Journal of the American Society for Information Science. 1989. 40(4): 284-290.

27. Baldi S. Normative versus Social Constructivist Processes in the Allocation of Citations: A Network-Analytic Model. American Sociological Review. 1998. 63(6): 829-846.

28. Merton RK. The Matthew Effect in science, II: cumulative advantage and the symbolism of intellectual property. ISIS. 1988. 79: 606-623.

29. Cronin B, *The hand of science. Academic writing and its rewards*. 2005: Scarecrow Press.

30. White H. Reward, persuasion, and the Sokal Hoax: A study in citation identities. Scientometrics. 2004. 60(1): 93-120.

31. Aphinyanaphongs Y, Statnikov A, Aliferis CF. A comparison of citation metrics to machine learning filters for the identification of high quality MEDLINE documents. JAMIA. 2006. 13(4): 446-455.

32. Bharat K, Henzinger M. *Improved algorithms for topic distillation in a hyperlinked environment*. in *21st ACM International Conference on Research and Development in Information Retrieval*. 1998.

33. Borodin A, Roberts G, Rosenthal J, et al. *Finding authorities and hubs from links structures on the world wide web*. in *10th International World Wide Web Conference*. 2001.

34. Kleinberg J. Authoritative sources in a hyperlinked environment. Journal of the ACM. 1999. 46(5): 604-632.

35. Haveliwala T. Topic-sensitive PageRank. IEEE Transactions on Knowledge and Data Engineering. 2003. 15(4): 784-796.

36. Nie L, Davison B, Qi X. *Topical link analysis for web search*. in *29th ACM Internation Conference on Research and Development in Information Retrieval*. 2006.

37. Richardson M, Domingos P. The intelligent surfer: probablistic combination of link and content information in PageRank. Advances in Neural Information Processing Systems. 2002.

38. Garfield E. Citation analysis as a tool in journal evaluation. Science. 1972. 178: 471-479.

39. Glanzel W, Moed H. Journal impact measures in bibliometric research. Scientometrics. 2002. 53(2): 171-193.

40. Takahashi K, Aw T, Koh D. An alternative to journal-based impact factors. Occup Med. 1999. 49(1): 57-59.

41. Uehara M, Takahashi K, Hoshuyama T, et al. A proposal for topic-based impact factors and their application to occupational health literature. J Occup Health. 2003. 45(4): 248-253.

42. MeSH Browser: National Library of Medicine. http://www.nlm.nih.gov/mesh/MBrowser.html. (accessed Aug 2008).

43. Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986. i: 307-310.

44. ISI Web of Science: Thomson Scientific. http://www.isiknowledge.com (accessed Aug 2008).

45. PubMed Clinical Queries Table: National Library of Medicine. http://www.ncbi.nlm.nih.gov/entrez/query/static/clinicaltable.html. (accessed August 2008).

46. ACP Journal Club: American College of Physicians. http://www.acpjc.org. (accessed August 2008).

47. Lempel R, Moran S. Rank-stability and rank-similarity of link-based web ranking algorithms in authority-connected graphs. Information Retrieval. 2005. 8: 245-264.

48. Ng A, Zhen A, Jordan M. *Stable algorithms for link analysis*. in *24th ACM International Conference on Research and Development in Information Retrieval*. 2001.

49. Kamvar S, Haveliwala T, Manning C, et al. Exploiting the block structure of the web for computing PageRank. Stanford University Technical Report. 2003.

50. Stanford WebBase Project: Stanford University. http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase. (accessed December 2007).

51. Feitelson D, Yovel U. Predictive ranking of computer scientists using CiteSeer data. Journal of Documentation. 2004. 60(1): 44-61.

52. Getoor L. Link mining: a new data mining challenge. SIGKDD Explorations. 2003. 5(1): 84-89.

53. Redner S. Citation statistics from 110 years of Physical Review. Physics Today. 2005. 58(6): 49-54.

54. Rattigan M, Jensen D. The case for anomalous link discovery. SIGKDD Explorations. 2003. 5(1): 41-47.

55. Hudson J. Be known by the company you keep: citations - quality or chance? Scientometrics. 2007. 71(2): 213-238.

56. Gehrke J, Ginsparg P, Kleinberg J. Overview of the 2003 KDD CUP. SIGKDD Explorations. 2003. 5(2): 149-151.

57. Manjunatha J, Sivaramakrishnan K, Raghavendra K, et al. Prediction using time series approach KDD Cup 2003 (Task 1). SIGKDD Explorations. 2003. 5(2): 152-153.

58. Csárdi G. Dynamics of Citation Networks. Proceedings of the International Conference on Artificial Neural Networks. 2006. 698-709.

59. Castillo C, Donato D, Gionis A. Estimating the number of citations using author reputation. Proceedings of String Processing and Information Retrieval (SPIRE). 2007. 107-117.

60. Lokker C, McKibbon KA, McKinlay RJ, et al. Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. BMJ. 2008. http://www.bmj.com/cgi/content/abstract/bmj.39482.526713.BEv1.

61. Popescul A, Ungar L. Statistical Relational Learning for Link Prediction. Proceedings of the Workshop on Learning Statistical Models from Relational Data at IJCAI 2003. 2003.

62. Taskar B, Wong M, Abbeel P, et al. Link prediction in relational data. Advances in Neural Information Processing Systems (NIPS 2003). 2004.

63. Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. J Am Soc Info Sci Technol. 2007. 58(7): 1019-1031.

64. Al-Hasan M. Link prediction using supervised learning. SIAM Workshop on Link Analysis, Counterterrorism and Security with SIAM Data Mining Conference. 2006.

65. Academic Ranking of World Universities: Shanghai Jiao Tong University. http://ed.sjtu.edu.cn/ranking2006.htm (accessed Aug 2008).

66. ISI Highly Cited Researchers: Thomson Scientific. http://www.isihighlycited.com/. (accessed Aug 2008).

67. Stopwords: National Library of Medicine. http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.table.pubmedhelp.T43. (accessed Aug 2008).

68. Porter M. An algorithm for suffix stripping. Program. 1980. 14: 130-137.

69. Leopold E, Kindermann J. Text categorization with support vector machines. Machine Learning. 2002. 46: 423-444.

70. Burges C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 1998. 2(2): 121-167.

71. Muller K, Mika S, Ratsch G, et al. An introduction to kernel-based learning algorithms. IEEE Trans. on Neural Networks. 2001. 12(2): 181-201.

72. Aliferis CF, Statnikov A, Tsamardinos I, et al. Local Causal and Markov Blanket Induction Algorithms for Causal Discovery and Feature Selection for Classification. Part I: Algorithms and Empirical Evaluation. Accepted to Journal of Machine Learning.

73. Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. Machine Learning. 2002. 46: 389-422.

74. LIBSVM -- A Library for Support Vector Machines: Chang C, Lin C. http://www.csie.ntu.edu.tw/~cjlin/libsvm/. (accessed Mar 2008).

75. Aliferis CF, Statnikov A, Tsamardinos I. Challenges in the Analysis of Mass-Throughput Data. Cancer Informatics. 2006. 2: 133-162.

76. Glenisson P, Glanzel W, Persson O. Combining full-text analysis and bibliometric indicators. A pilot study. Scientometrics. 2005. 63(1): 163-180.

77. Meho L, Kiduk Y. Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. J Am Soc Info Sci Technol. 2007. 58(13): 2105-2125.

78. Brody T, Harnad S, Carr L. Earlier web usage statistics as predictors of later citation impact. J Am Soc Info Sci Technol. 2006. 57(8): 1060-1072.

79. Garzone M, Mercer RE. *Towards an automated citation classifier*. in *Canadian Conference on AI*. 2000.

80. Mercer RE, DiMarco C. *A design methodology for a biomedical literature indexing tool using the rhetoric of science*. in *2004 Joint Conference on Human Language Technology/North American Association for Computational Linguistics (HLT-NAACL)*. 2004. Boston, MA.

81. Teufel S, Siddharthan A, Tidhar D. *Automatic classification of citation function*. in *Proceedings of EMNLP*. 2006. Sydney, Australia.

82. Egashira K, Inou T, Hirooka Y, et al. Evidence of Impaired Endothelium-Dependent Coronary Vasodilatation in Patients with Angina Pectoris and Normal Coronary Angiograms. New England Journal of Medicine. 1993. 328(23): 1659-1664.

83. Hecht SS, Carmella SG, Murphy SE, et al. A Tobacco-Specific Lung Carcinogen in the Urine of Men Exposed to Cigarette Smoke. New England Journal of Medicine. 1993. 329(21): 1543-1546.

84. Siscovick DS, Raghunathan TE, Psaty BM, et al. Diuretic Therapy for Hypertension and the Risk of Primary Cardiac Arrest. New England Journal of Medicine. 1994. 330(26): 1852-1857.