Journal of Biomedical Informatics 44 (2011) 587–594

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A comparison of evaluation metrics for biomedical journals, articles, and websites in terms of sensitivity to topic

Lawrence D. Fu [a,*], Yindalon Aphinyanaphongs [a,1], Lily Wang [b,2], Constantin F. Aliferis [a,1]

[a] New York University Medical Center, Center for Health Informatics and Bioinformatics, 227 E 30th St., 7th Floor, New York, NY 10016, USA
[b] Department of Biostatistics, Vanderbilt University School of Medicine, S-2323 Medical Center North, Nashville, TN 37232, USA

ARTICLE INFO

Article history:
Received 1 June 2010
Available online 17 March 2011

Keywords:
Information retrieval
Machine learning
PageRank
Journal impact factor
Topic-sensitivity
Bibliometrics

ABSTRACT

Evaluating the biomedical literature and health-related websites for quality are challenging information retrieval tasks. Current commonly used methods include impact factor for journals, PubMed's clinical query filters and machine learning-based filter models for articles, and PageRank for websites. Previous work has focused on the average performance of these methods without considering the topic, and it is unknown how performance varies for specific topics or focused searches. Clinicians, researchers, and users should be aware when expected performance is not achieved for specific topics. The present work analyzes the behavior of these methods for a variety of topics. Impact factor, clinical query filters, and PageRank vary widely across different topics while a topic-specific impact factor and machine learning-based filter models are more stable. The results demonstrate that a method may perform excellently on average but struggle when used on a number of narrower topics. Topic-adjusted metrics and other topic robust methods have an advantage in such situations. Users of traditional topic-sensitive metrics should be aware of their limitations.

1. Introduction

The size of the biomedical literature and 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 becoming increasingly resource prohibitive or otherwise impractical. Such methods include the impact factor to measure journal impact or quality [1]. PubMed clinical query filters measure methodological quality of papers in well-defined content categories [2]. 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 [3,4]. Finally, the most popular way to rank the quality of web pages is PageRank [5].

Previous studies have measured the performance of these methods *across a wide range of topics* although clinicians, researchers, and users typically search for specific topics each time. The variability of these methods for different topics is currently unknown, and it is possible for a method with excellent average performance to fail in focused domains. To make the point clear, suppose we have a set of articles about two topics (topic A and topic 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. The problem is that a researcher interested in only topic B would unknowingly experience much worse than expected performance.

In web-related research, topic sensitivity is known as *topic drift* where the topic of the top results is different from the query topic [6]. PageRank is known to suffer from topic drift and may not yield the best results for a specific topic. A simple example that shows the possibility is the following: suppose we have two web pages with different degrees of relevance to a topic of interest (topic A). If the first page receives most of its links from pages related to another topic (topic B), it is only marginally relevant to topic A. If the second page receives most of its links from pages about topic A, it is more relevant to topic A. However, PageRank is calculated in a manner where the unrelated first page may be ranked higher than the second page although the second page is a better resource for the topic of interest. The first page could be cited by many pages which are of high quality but related to a different topic.

The purpose of this work is to examine the extent to which performance varies with different topics for journal impact factor, clinical query filters, machine learning pattern recognition methods, and PageRank when identifying high quality journals, articles,

* Corresponding author. Fax: +1 212 263 5995.
E-mail addresses: lawrence.fu@nyumc.org (L.D. Fu), yin.a@nyumc.org (Y. Aphinyanaphongs), lily.wang@vanderbilt.edu (L. Wang), constantin.aliferis@nyumc.org (C.F. Aliferis).
[1] Fax: +1 212 263 5995.
[2] Fax: +1 615 343 4924.

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.03.006

and websites. We also examine modified approaches that adjust for topic or are insensitive to topic as alternatives to highly topic-sensitive methods.

2. Methods

We utilize tailored methods for evaluating journal, articles, and websites respectively. For each document type, there are three sections: methods, experimental design, and corpus or experimental data set construction.

2.1. Evaluation methods for journals

2.1.1. Journal impact factor

The journal impact factor evaluates journal impact regardless of publication size or frequency [1,7,8]. It affects journal readership and helps researchers determine which journal to 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 [1]:

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

The numerator is the number of citations received in a given year to journal items published in the previous 2 years. The denominator is the number of journal articles from the previous 2 years. The numerator includes articles, editorials, and letters to the editor, while the denominator consists only of articles. 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 to 2003 divided by the number of articles from 2002 to 2003.

2.1.2. 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) [9,10]. However, this metric does not assess journals. We need a formula that isolates the contribution of a specific topic from the overall impact factor to study the sensitivity of impact factor to topic. We calculate a topic-specific impact factor (TIF) 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 to 2003. The denominator is the number of cardiology-related articles. Determining topic relevance is topic-specific. For example, we 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 [11]. When we specify topics, the topics do not need to be exclusive or cover all items for the adjustment to be meaningful.

2.1.3. 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. We define the topic-mix adjusted impact factor for k topics as:

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

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.

2.2. Evaluation methods for articles

2.2.1. Clinical query filters

The clinical query filters were originally designed by Haynes and colleagues [2] and are the most widely available method for identifying high-quality articles through PubMed [2]. These filters are semi-manually constructed Boolean queries of terms in the MeSH headings, publication type, or text of the MEDLINE record. All articles that match a given combination of terms are returned. Performance of these filters is typically measured by sensitivity and specificity. Filters are defined for diagnosis, etiology, prognosis, and treatment with queries optimized for sensitivity and specificity. 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.

2.2.2. Support vector machine models

Machine learning methods are another approach to identifying high-quality articles. In previous research, polynomial support vector machine models [12] had superior performance compared to clinical query filters [3,4]. These models preprocess fields and text from MEDLINE records by converting them into input features for learning. A kernel function maps the input space to a feature space where a hyperplane is calculated to separate the classes of data. We used the models learned from a previous study [3] which includes further details about the learning procedure. Performance is measured by area under the receiver operating curve (AUC).

2.3. Evaluation method for websites

2.3.1. PageRank

PageRank is a citation-based method for evaluating the quality of web pages [5]. One way to explain this method is that it considers a page to be of high quality if many pages link to it and these pages are also of high quality using the same criterion recursively. A more mathematical way to describe PageRank is that it models user behavior as a random surfer who ignores page content by either arbitrarily following a link or randomly jumping to a page. The PageRank of a page is proportional to the likelihood that the random surfer visits a page. A page with a high PageRank will be linked by many pages or by pages with high PageRanks.

The PageRank of a page u is then calculated as follows:

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

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 specifies the probability of following

a link. The surfer will jump to a random page with probability $1 - \alpha$ or follow an outlink with probability α. It is typically set at .85 but can take any value between 0 and 1. The first term in the equation is the probability of visiting a page after a random jump. The second term is the sum of all PageRanks from all incoming links. For each inlink, the PageRank is divided by the number of links for that page. These values are summed over all incoming links and weighted by α. A vector of PageRank values is defined for all pages in the network, and PageRank calculations are performed as matrix operations. The PageRank values are guaranteed to converge by adding links to pages without any links and having the random jumps.

Researchers have modified PageRank to address the issue of topic drift. Haveliwala computed topic-sensitive PageRank scores by calculating a score for each page with respect to a number of topics [13]. The topics were top level categories from the Open Directory Project. The topic-sensitive PageRanks were computed by biasing the random jump step to favor pages related to a given topic. The final PageRank values were computed at query time by weighting each topic-sensitive PageRank according to how similar the topic was to the query. Richardson and Domingos [6] used an intelligent surfer model which analyzed 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 et al. [14] augmented the random surfer model by using a topical random surfer that considered topics while surfing. When a surfer follows an outlink, the surfer could stay on the same topic or change the topic of interest.

3. Results

3.1. Results for journal methods

Experiments for the journal methods were performed on a defined 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 represent a wide range of topics and citability. We selected eight general topics of internal medicine and a set of narrowly-defined random subtopics from Gastroenterology. All topics were used as defined by the MeSH vocabulary. For each journal and topic, all relevant MEDLINE records from 2003 to 2004 were retrieved, and citation counts and journal impact factors were obtained from the ISI Web of Knowledge [15].

We calculated the absolute differences of impact factor to topic-specific impact factor to analyze how much impact factors varied for different topics. There should be little difference between impact factor and topic-specific impact factor if impact factor is stable for different topics. The results in Table 1 show that a higher overall impact journal did not always have a higher topic-specific impact factor. For example, NEJM had a higher impact factor than JAMA but had a lower cardiology-specific impact factor. Of the 120 comparisons among the 15 journal pairs and 8 topics, there were 10 reversals (8.33% of the comparisons, 95% confidence interval 3.39–13.28%) where a higher impact journal had a lower topic-specific impact factor. There were three 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). The results show that rankings based on impact factor and topic-specific impact factor are not always equivalent.

The minimum, median, maximum, and interquartile ranges of the differences were calculated to assess the skewness and spread of the values. Interquartile range measures dispersion and is the difference between the third and first quartiles. When computing topic-specific impact factors, we do not have *p*-values or confidence intervals since they are population totals and not point estimates. Table 2 shows the values. The maximum differences ranged from about 10–35 which means that the values can vary greatly for different topics.

A Bland–Altman plot [16] determined whether topic-specific impact factors coincided with impact factors or if they were significantly different. This graph 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. We note that the correlation coefficient is not an optimal method for judging agreement among methods and thus was not pursued here [16]. The Bland–Altman plot in Fig. 1 shows that the difference between 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 specialty since all topics showed some difference. If impact factor and topic-specific impact factor were equivalent, all values would appear 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 Bland–Altman plot, along with the absolute differences between impact factor and topic-specific impact factor, demonstrate that the two methods are not always equivalent.

The observations for the eight general topics from internal medicine were also evident for gastroenterology subtopics (data not shown due to space restrictions). As with the general topics, there

Table 1
Topic-specific impact factors for general topics and journal impact factor in 2004 and 2003.

	Journal	Topic-specific impact factors for general topics								Impact factor
		Cardiology	Endocrinology	Gastroenterology	Hematology	Medical oncology	Nephrology	Pulmonary disease	Rheumatology	
2004	AIM	16.07	13.85	16.92	7.94	12.49	23.17	12.66	15.40	13.11
	AJM	4.09	3.44	2.73	6.38	3.95	4.31	3.10	6.29	4.18
	BMJ	7.55	6.48	7.37	5.73	5.57	2.37	7.94	8.77	7.04
	JAMA	42.18	28.27	60.55	13.87	35.58	20.32	36.47	13.40	24.83
	Lancet	33.80	47.70	18.86	11.98	23.16	14.30	27.41	52.50	21.71
	NEJM	37.46	54.31	37.68	33.71	44.8	27.93	37.97	24.08	38.57
2003	AIM	14.37	19.83	12.73	10.63	12.14	23.06	13.21	14.50	12.43
	AJM	4.21	5.82	2.43	4.30	3.98	5.33	3.44	5.82	4.40
	BMJ	7.95	6.84	4.98	5.57	5.76	4.00	5.37	12.25	7.21
	JAMA	38.12	28.24	70.00	13.38	39.27	18.94	30.13	12.80	21.46
	Lancet	24.42	34.33	17.91	8.34	17.78	14.61	14.12	17.94	18.32
	NEJM	38.05	55.78	33.66	28.78	40.46	39.51	22.42	45.33	34.84

Table 2
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	15.00
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	5.10
Rheumatology	1.73	6.86	30.79	12.38

were a number of ranking reversals. The variation increased for more specialized topics and was most pronounced in the three highest impact journals. JAMA had the greatest variability with a maximum topic-specific impact factor that was over 13 times larger than its minimum. The overall impact factor became less meaningful for increasingly specialized topics. 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 clearly suggest that researchers studying a specific disease should not rely on overall impact factor for journal evaluation.*

We performed additional experiments to ensure that variation was not a random occurrence unique to a single year. First, we replicated the experiments for 2003 and found consistent results as shown in Table 1. Many of the relative rankings of the journals were consistent, while some of the same reversals existed. Ranges of topic-specific impact factors were also comparable. Next, we verified that variation was not randomly caused by smaller sample sizes independent of topic. By definition, journal impact factor is calculated on a larger number of publications than the topic-specific impact factor. We tested whether the difference between the two measures was associated with sample size by computing the regression coefficients of the following regression model:

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

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. We found that β_1, the regression coefficient for sample size difference, 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.

For an example of a topic-mix adjusted impact factor, we used the 2004 data and a topic mix where cardiology is 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 either.

3.2. Results for article methods

Experiments for the article methods were performed on a corpus previously used to compare clinical query filters and SVM models [3]. The ACP Journal Club [17] was used as the gold standard. It is a meta-publication where experts review the best journals in internal medicine monthly to identify high-quality articles for categories including 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 journal articles from the same period were negative cases or not considered high-quality. There were 15,786 MEDLINE records from July 1998 to August 1999 for the treatment and etiology categories. There were 34,938 MEDLINE records from July 1998 to August 2000 for prognosis and diagnosis. The longer timeline enabled the collection of a sufficient number of positive cases. Articles were formatted for learning by extracting and encoding terms from the abstract, title, MeSH terms, and publication type.

We measured the topic-sensitivity of the clinical query filters and the SVM models by observing the change in performance when articles were separated by topics. Overall performance was first calculated for all articles. The performance metrics were sensitivity and specificity for clinical query filters, and the metric was AUC for the SVM models. Then, performance was measured for subsets of articles related to a specific topic. We randomly selected 18 MeSH terms covering a range of topics. The topics were: Bone Diseases, Cardiovascular Diseases, Cysts, Diabetes Mellitus, Endocrine System Diseases, Gastroenteritis, Gastrointestinal Diseases,

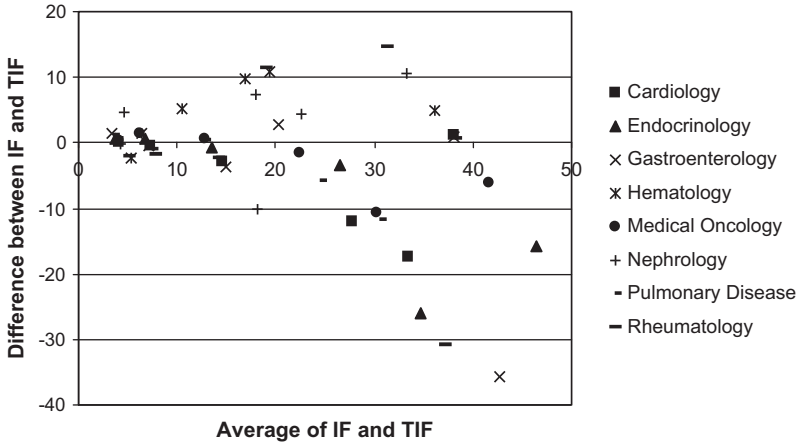

Fig. 1. Bland–Altman plot for differences between impact factor and topic-specific impact factor.

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.

Absolute differences were computed between the overall performance and the topic-specific performances. We computed the minimum, median, maximum, and interquartile ranges of these differences to summarize the results. We also performed Wilcoxon signed rank tests which test the difference between paired measurements. 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 maintain the same performance for individual topics.

Table 3 summarizes the differences between the overall sensitivity/specificity and the observed values for clinical query filters. Appendix A shows the actual values for specific topics. There was considerable variability for clinical query filters within 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 sensitivity/specificity for the sensitivity-optimized prognosis filter. A p-value less than .05 means that the method does not maintain the same performance for individual topics.

The SVM models were more stable over topics as shown in Table 4. Appendix B shows the actual values for specific topics. The results cannot be compared directly with the Haynes' filters results since AUC values are not sensitivity or specificity values. However, AUC values also range from 0 to 1. Differences are much smaller since the largest interquartile range is .065, and the largest maximum difference is 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 machine learning models are less sensitive to topic and are 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 four positive cases out of several 100 or 1000 articles. The diagnosis results may be consistent with the results for the other categories if given more positive cases.

3.3. Results for websites

3.3.1. Experimental considerations for websites

Experiments for PageRank were performed on a collection of web pages provided by the Stanford WebBase [22] which is a repository of topic-focused web crawls intended for research use.

Table 4
The minimum, median, maximum, and interquartile ranges for the absolute differences between overall and topic-specific AUC values for SVM models.

Category	Min	Median	Max	IQR
Diagnosis	0.0083	0.038	0.04	0.012
Etiology	0.0027	0.028	0.13	0.050
Prognosis	0.0041	0.045	0.10	0.065
Treatment	0.00054	0.0040	0.041	0.0078

WebBase provided link and html information, but only the link structure was needed for our purposes.

A number of technical issues had to be resolved before the topic-sensitivity of PageRank could be evaluated. Research involving the web 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 all incoming links to a page have been detected. Researchers typically create a static snapshot of the web by sampling pages. PageRank values are affected when pages are removed since the network topology changes dramatically as links are removed, and it is not completely understood how sampling affects the stability of rankings [18–20].

Given these considerations, we made two decisions about how to sample web pages. First, we sampled networks by selecting pages from the same domain. Kamvar demonstrated that most pages link to pages from the same domain [21]. He found that 83.9% of links connected pages from the same domain in his test corpus. The percentage rose to 95.2% after removing pages without outlinks. Sampling pages from the same domain appears to minimize the effect on PageRank.

The second sampling decision was to select high-ranking pages. Ng [20] showed that removing pages with low PageRank does not affect the stability of the top 10 results. We investigated whether rankings are stable for all results since users may be interested in more than 10 results. PageRanks were first computed for all pages within a domain. Then the lowest ranking pages were removed, and PageRanks were computed for the remaining pages. The stability of rankings was calculated using Haveliwala's Ksim metric [13], which is based on Kendall's τ distance measure. Ksim is the fraction of pairwise ranking comparisons that are consistent between both sets of rankings. If page A is ranked higher than page B in one set of rankings, it verifies whether 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 steps of removing pages and re-calculating PageRanks were repeated until a small number of pages remained.

Four domains were chosen: the National Diabetes Education Program (ndep.nih.gov, 415 pages), the National Eye Institute (www.nei.nih.gov, 1151 pages), the National Heart Lung and Blood Institute (www.nhlbi.nih.gov, 3784 pages), and the Centers for Disease Control and Prevention (www.cdc.gov, 9434 pages). These do-

Table 3
The minimum, median, maximum, and interquartile ranges of the absolute differences between overall and topic-specific sensitivity/specificity for clinical query filters. The specificity-optimized filter for diagnosis did not return any articles.

Optimized for	Category	Sensitivity				Specificity			
		Min	Median	Max	IQR	Min	Median	Max	IQR
Sensitivity	Diagnosis	0.020	0.02	0.15	0.0013	0.015	0.087	0.23	0.10
	Etiology	0.028	0.07	0.070	0.00	0.00047	0.059	0.22	0.10
	Prognosis	0.031	0.10	0.57	0.15	0.0029	0.053	0.18	0.04
	Treatment	0.004	0.010	0.030	0.0025	0.0027	0.03	0.17	0.05
Specificity	Diagnosis	–	–	–	–	–	–	–	–
	Etiology	0.16	0.34	0.49	0.28	0.0066	0.13	0.31	0.09
	Prognosis	0.11	0.24	0.52	0.33	0.030	0.099	0.22	0.04
	Treatment	0.034	0.053	0.070	0.023	0.00037	0.048	0.13	0.030

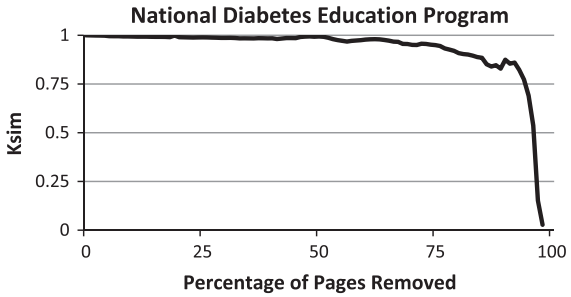

Fig. 2. Ksim values for subsets of NDEP as pages removed (initially 415 pages).

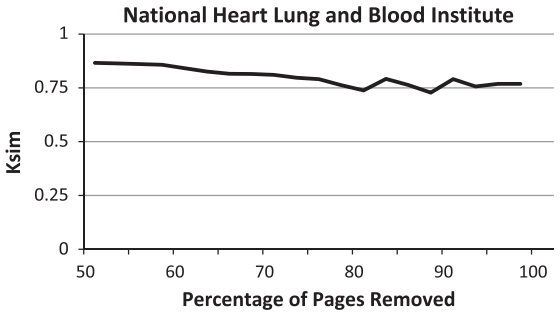

Fig. 5. Ksim values for subsets of NHLBI as pages removed (initially 3784 pages).

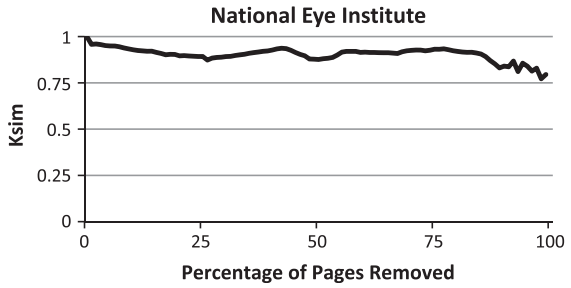

Fig. 3. Ksim values for subsets of NEI as pages removed (initially 1151 pages).

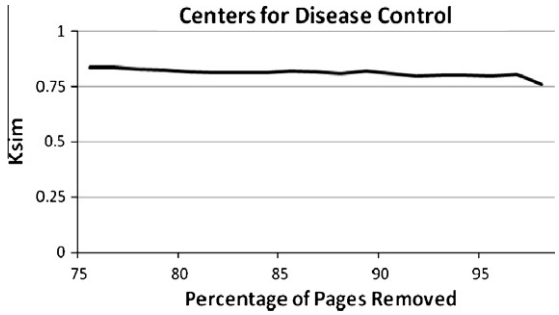

Fig. 4. Ksim values for subsets of CDC as pages removed (initially 9434 pages).

mains were selected to provide biomedically relevant samples of various sizes. The stability of rankings was measured by repeatedly removing pages with low PageRanks and re-computing PageRanks. Experiments for the CDC and NHLBI domains were started by removing the lowest ranked 25% and 50% of pages respectively (shown in Figs. 4 and 5). These domains contain a larger number of sites compared to the NDEP and NEI domains. The running time for computing Ksim values becomes prohibitive for large domains because of the large number of pairwise comparisons required.

Figs. 2–5 show that rankings do not fluctuate dramatically if low ranking pages are removed. All domains had Ksim values over 0.8 after the first removal. The Ksim values gradually decreased with fewer pages until a small number remained. The results indicate that sampling a network by selecting high-ranking pages is a reasonable method for creating a subset with consistent rankings.

3.3.2. Studying the topic-sensitivity of PageRank

After deciding how to sample web pages while minimizing the effect on PageRank, we measured the variability of PageRank for different topics by removing pages unrelated to a given topic. It is possible for highly-ranked pages in a network with a mixture of topics to receive many links from pages unrelated to a topic of interest. These pages could decrease in rank within a topic-specific network. The analysis involved first computing PageRanks for all pages. Then, we isolated the pages related to a specific topic, recomputed PageRanks on this subset, and assessed the similarity between the two sets of rankings. High-ranking pages in the original network could decrease in rank in the topic-isolated network if it received many links from unrelated pages.

Evaluating the similarity is not as straightforward as simply applying the Ksim metric. Sampling topic-specific networks drastically changes the network topology by removing links which affects PageRank values. Similarity differences are then caused by random fluctuations from the changing topology as well as the effect of topic isolation. Distinguishing between these two causes was necessary to accurately measure topic-sensitivity. To address this problem, we generated random subsets of the same size as each topic-isolated subset, and PageRanks were computed for the random subset. The similarity between the original network and random subsets was measured with Ksim, and results were averaged over 5 runs. The Ksim values for the random subsets provided a baseline for comparison. Ksim values were then computed between the original network and the topic-isolated subsets. These values were compared to the values for the random subsets, and any increase was attributed to the effect of isolating for topic.

Two health-related domains were used: the Centers for Disease Control and Prevention (www.cdc.gov, 9434 pages) and the National Cancer Institute (www.cancer.gov, 9708 pages). A number of well-represented topics were selected for each domain. The CDC site was organized in a directory structure that prompted us to separate pages according to the following categories: 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 domain, a website was considered relevant to a topic if the address contained a related word. For example, a page was included in the "lung" topic if "lung" or "pulm" was in the address. The topics used were breast, cervix, colon, lung, and prostate.

Table 5 displays the Ksim values for the two domains. Ksim values for the topic subsets were greater than the values for the random subsets which implies that the rankings are dependent on topic. It is important to note that the Ksim values for the CDC topics were higher than the NCI topics. The CDC rankings were more stable because of the prevalence of intra-topic linking in the original network. At least 71% of the links in the original CDC network originated from pages related to each topic. Removing unrelated pages does not considerably alter the rankings if pages mostly link to pages within the same topic. On the other hand, rankings are very unstable after removing unrelated pages if most links connect pages unrelated to the topic. The Tobacco and Genomics topics in the CDC site had the greatest proportion of intra-topic links (94% and 97%) and the highest Ksim values (.87 and .85). The Breast and Prostate topics in the NCI site had the lowest proportions of

Table 5
Ksim values for topic-isolated and random subsets along with the percentage of links remaining after topic isolation.

Domain	Topic	Num. pages	Ksim for topic subset	Ksim for random subset	Proportion of links within topic
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
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

intra-topic links (32% and 32%) and two of the lowest Ksim values (.71 and .7). The correlation between the two measurements shows that the stability of PageRank for different topics depends on the proportion of links from pages related to the topic of interest. In other words, PageRank is more stable with a greater proportion of related links.

4. Discussion

The present paper studied the variability of evaluation metrics for the scientific literature and web when considering different topics. For journals and articles, previous research had studied the average performance of impact factor, clinical query filters, and SVM-based models over all topics. The present study analyzed the stability of these methods for specific topics as well as two aspects of PageRank's behavior for evaluating web pages. First, our results demonstrated that removing pages with low PageRanks is a reasonable sampling method since the rankings of the remaining pages were relatively stable. Second, the current study showed that rankings based on PageRank vary according to topic in direct correspondence to the variation of links from pages related to different topics of interest.

We initially hypothesized that impact factor, clinical query filters, and PageRank are unstable for different topics since they are query-independent methods built separately from the learning task. This was verified by our data. Conversely query-dependent methods always consider the search topic and are constructed for a particular learning task. Machine learning filter models and topic-specific impact factor are two examples. We hypothesized that query-focused methods would be less variant across topics which was confirmed by our data.

Citation-based metrics such as impact factor and PageRank are query-dependent methods that are highly topic variant according to our data. It is relatively straightforward to understand why: an average document often discusses multiple topics and receives multiple citations. However, all topics are not relevant to any one citation. This is compounded by the fact that a citation plays a variety of roles that do not necessarily constitute endorsement related to all the topics of interest for a document. An article may cite another article for a variety of reasons: to acknowledge prior work, identify methodology, correct or criticize, or disclaim the work of others [23].

We postulate without proof that the topic-sensitivity of clinical query filters is due to the manual, expert-driven process by which they are constructed. Experts choose terms that reflect their expertise in particular areas. The coverage of terms therefore may not be exhaustive since research areas use different jargon and vocabulary, and topics outside the experts' knowledge may lack adequate consideration. On the other hand, machine learning models automatically learn terms for all topics in the corpus. The machine learning methods in principle and in our experiments perform well for any topics that are sufficiently sampled in the corpus.

One factor that needs to be considered is that the set of papers and web pages is finite. Topic-adjusted metrics cannot be estimated with acceptable statistical certainty for very narrow topics. The issue of adequate sample size (or equivalently adequate statistical certainty about estimates for such metrics) should be considered whenever topic adjustments are made. Simple and standard power-size analysis and statistical inference techniques from statistics are adequate for this purpose [24].

Finally, one word of caution is warranted regarding the development cost of machine learning models. The primary costs for such models are the collection and construction of training corpora. This cost is expected to be amortized over several years worth of use by potentially thousands of users and models. Thus it is reasonable to expect very small development costs for popular literature and web searches. The machine learning approach however may be too costly to implement for some important yet infrequent searches where training corpora are hard to come by.

4.1. Conclusions and practical implications of the present work

The experiments and results discussed in the present paper point to specific, practical ways to improve current practices in information retrieval and academic quality assessment. Since the impact factor and clinical query filters are considerably sensitive to topic, we believe that they should be replaced eventually by versions that are adjusted for topic (subject to technical feasibility and other practical considerations). We presented specific ways to adjust the impact factor. We note however that the adjustment technique we proposed is fine grained (i.e., performed at the topic level, not the average of a field that contains many topics) but may be resource intensive to compute. Machine learning models do not exhibit topic sensitivity, so the practical implication is that the methodologies employed for constructing such models are sound and do not need adjustments (like citation-based methods do).

Until the topic adjusted approach to citation-based methods is widespread, users of uncorrected metrics will benefit by recognizing the limitations of traditional metrics. For example, researchers interested in gastrointestinal diseases would believe that the New England Journal of Medicine is the best journal to read according to impact factor. However, JAMA may be a better choice since it has higher topic-specific impact factors for gastroenterology and gastrointestinal diseases. We believe that there is an urgent need for topic-specific impact factor and citation count databases to allow for a more robust guidance for end users. Similar conclusions and practical implications apply to the realm of web page search and review. Our results suggest that next generation search engines will benefit by offering topic-adjusted search results ranking (at least for a core set of frequent queries or medical topic categories). Finally machine learning scoring of web pages with existing methods offers an attractive alternative technology for ranking web pages as long as machine learning models are built and made available to users for a class of widely used cases.

Acknowledgment

The authors gratefully acknowledge support from Grants R56 LM007948-04A1 and 1UL1RR029893.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2011.03.006.

References

[1] Garfield E. The history and meaning of the journal impact factor. JAMA 2006;295:90–3.
[2] Haynes R, Wilczynski N, McKibbon K, et al. J Am Med Inf Assoc 1994;1:447–58.
[3] Aphinyanaphongs Y, Tsamardinos I, Statnikov A, et al. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inf Assoc 2005;12:207–16.
[4] Aphinyanaphongs Y, Aliferis CF. Text categorization models for retrieval of high quality articles in internal medicine. AMIA annual symposium; 2003.
[5] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. In: Proceedings of the 7th international world wide web conference; 1998.
[6] Richardson M, Domingos P. The intelligent surfer: probabilistic combination of link and content information in PageRank. Adv Neural Inf Process Syst 2002;14:1441–8.
[7] Garfield E. Citation analysis as a tool in journal evaluation. Science 1972;178:471–9.
[8] Glanzel W, Moed H. Journal impact measures in bibliometric research. Scientometrics 2002;53:171–93.
[9] Takahashi K, Aw T, Koh D. An alternative to journal-based impact factors. Occup Med 1999;49:57–9.
[10] 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;45:248–53.
[11] National library of medicine. MeSH browser. <http://www.nlm.nih.gov/mesh/MBrowser.html> [accessed October 2010].
[12] Burges C. A tutorial on support vector machines for pattern recognition. Data Mining Knowledge Discov 1998;2:121–67.
[13] Haveliwala T. Topic-sensitive PageRank. IEEE Trans Knowledge Data Eng 2003;15:784–96.
[14] Nie L, Davison B, Qi X. Topical link analysis for web search. In: 29th ACM international conference on research and development in information retrieval; 2006.
[15] Thomson scientific. ISI web of knowledge. <http://www.isiknowledge.com> [accessed October 2010].
[16] Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet i 1986:307–10.
[17] American College of Physicians. ACP journal club. <http://www.acpjc.org> [accessed May 2010].
[18] 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.
[19] Lempel R, Moran S. Rank-stability and rank-similarity of link-based web ranking algorithms in authority-connected graphs. Inf Retrieval 2005;8:245–64.
[20] Ng A, Zhen A, Jordan M. Stable algorithms for link analysis. In: 24th ACM international conference on research and development in information retrieval; 2001.
[21] Kamvar S, Haveliwala T, Manning C, et al. Exploiting the block structure of the web for computing PageRank. Stanford University Technical Report; 2003.
[22] Stanford webbase project. <http://diglib.stanford.edu:8091/~testbed/doc2/WebBase/> [accessed October 2010].
[23] Garfield E. Can citation indexing be automated? National Bureau of Standards; 1965.
[24] Tabachnick BG, Fidell LS. Using multivariate statistics. 5th ed. Boston: Allyn & Bacon; 2006.