Information Sciences 214 (2012) 76–90

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Combining relevancy and methodological quality into a single ranking for evidence-based medicine

Sungbin Choi [a], Borim Ryu [a], Sooyoung Yoo [b], Jinwook Choi [a,*]

[a] *Seoul National University College of Medicine, Seoul, Republic of Korea*
[b] *Seoul National University Bundang Hospital, Gyeonggi-do, Republic of Korea*

ARTICLE INFO

Article history:
Received 14 September 2011
Received in revised form 25 May 2012
Accepted 31 May 2012
Available online 9 June 2012

Keywords:
Evidence-based medicine
Ranking
Classification
Document quality

ABSTRACT

Evidence-based medicine has recently received a large amount of attention in medical research. To help clinical practices use evidence-based medicine, it should be easy to find the best current evidence that is relevant to the clinical question and has high methodological quality. However, searching for relevant articles and appraising their validity is demanding work for most clinicians. We hypothesize that, through an effective design that addresses the two major aspects - relevance and quality - together with a ranking algorithm, search engines can automatically retrieve articles that are relevant to clinical questions and are based on valid evidence. The contribution of this study has two parts. First, we approach this problem by combining methodologies. After designing a suitable document query-relevance score and methodological quality score, we combined them using various fusion methods. The result was a twofold increase in the mean average precision. Second, for correct evaluation, we built a test collection using a preexisting reliable database, the *Cochrane Reviews*, which allowed robust and comprehensive evaluation.

1. Introduction

Since its release in November 2004, *Google Scholar* [6] has gained much popularity among researchers and students in many fields, including medicine. Among the many features of *Google Scholar*, convenience and efficiency offer the largest advantages. By typing only keywords, we can quickly obtain relevance-ranked search results. Compared to a Boolean search engine such as *PubMed* [9], we do not have to agonize over constructing appropriate Boolean query combinations or spend time finding relevant articles from the retrieved documents. Although its thoroughness in searching might not reach the level of Boolean query strategies that an expert manually crafts, *Google Scholar* returns the most important and highly cited articles for even a non-professional searcher. *Google Scholar* describes its aim as ’Ranking documents the way researchers do’ [41]. This type of intelligence-flavored smart ranking algorithm can be useful, especially for clinicians who attempt to apply evidence-based medicine (EBM) in their daily practice.

EBM is widely recognized as an important concept in medical research. Evidence-based health care is the conscientious use of the best current evidence to make decisions about patient care or delivering health services. The best current evidence is up-to-date information from relevant, valid research about the effects of different forms of health care [21]. In [14], Ghosh et al. state that the future competence of a physician is not measured by his or her ability to recall facts but by his or her ability to incorporate the best current evidence into the patient’s personal values.

[*] Corresponding author.
E-mail address: jinchoi@snu.ac.kr (J. Choi).

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.05.027

However, practicing EBM in daily clinical care may be challenging, considering a physician's limited time and possibly inadequate searching skills [16]. EBM includes an appraising step, critically evaluating an article's evidence to decide whether it is reliable and robust [14]. Searching for relevant articles and assessing their validity is a demanding task.

We approached this problem by regarding it as an information retrieval task with two distinct priorities: finding sufficient research articles relevant to the clinician's question and finding valid articles based on EBM methodological criteria and principles. We hypothesize that a search engine designed to consider those two aspects together can retrieve articles that are relevant and valid. Using various fusion algorithms, we combined the relevant feature and methodological quality scores into a single ranking.

In this paper, we first built a test collection (Section 3.2) using preexisting sources (*Cochrane Reviews*). Second, we used a probabilistic retrieval model and machine learning classifier to determine each document's query relevancy and quality scores, respectively (Sections 3.3.1 and 3.3.2). We applied various fusion techniques to re-rank the retrieved documents (Section 3.3.3).

2. Background

2.1. PubMed and Google Scholar

PubMed is a free database that accesses the MEDLINE database of citations, abstracts, and some full-text articles on life science and biomedical topics [9]. *PubMed* currently contains over 21 million publications and offers a comprehensive search over the biomedical literature with advanced search features.

However, some people find *PubMed* difficult to use. The effectiveness of a Boolean search engine depends entirely on the user, because constructing complex Boolean queries that narrow the retrieved set to mostly relevant documents requires considerable experience [20]. To use Boolean search engines properly, users must be well trained. Building an appropriate Boolean query and exhaustively exploring the search results can also be cumbersome, demanding many iterations of work. *Google Scholar*, conversely, provides a rather simple way of searching the literature. The search results are ranked in an attempt to capture articles '*The way researchers do*'. To truly capture the way that researchers work, a method must consider who wrote the paper, where it was published, how relevant the article was to the query, and what other articles have said about it [41]. However, *Google Scholar* has been criticized for not being designed for a comprehensive literature search, instead being a "plug-in-the-keyword-and-hope-for-the-best-tool" [76].

Both *PubMed* and *Google Scholar* are widely recognized search engines in medical research. Trying to make a direct performance comparison between *PubMed* and *Google Scholar* could be a faulty approach, as each has been built for different purposes and use cases.

In this paper, we tried to make the system take some part of ranking process intelligently, instead of executing direct orders with every work detail written by the user. We can thus say that our method is more closely positioned to *Google Scholar*'s approach.

2.2. Text classification

Much text classification research has been devoted to various tasks, including topical categorization [69] and spam detection [24]. Until the late 1980s, applications using manually defined, hand-crafted rules to encode expert knowledge about classifying documents were popular in the operational field [69]. With careful work, these systems could achieve remarkable accuracy, although they required much time to create and fine-tune those rules [57]. As the Internet has become widely available and the amount of digital documents is growing rapidly, automatic text classification techniques that learn from data gain more importance. They apply machine learning methods, including the decision tree [17], cognitive situation model [31], Naïve Bayes [26] or support vector machine (SVM) [47] approaches, to categorize documents into predefined classes. Previous studies present more extensive surveys of these techniques [69,78].

Recently, text classification has also been applied to other domain-specific tasks, such as sentiment analysis and medical document quality assessment. In a sentiment analysis task, researchers try to monitor favorable or unfavorable sentiments towards specific entities by utilizing available sources, including online reviews or personal blogs [59]. For a medical document quality assessment task, many studies have considered automatically identifying methodologically high-quality articles [16,23,34–36,50]. We consider each task below.

2.3. Medical document classification

Evidence quality. Researchers have examined evaluating the information quality of web documents based on linguistic features [38,39]. Those works assess web document quality based on technical (e.g., information ordering, navigation) and content requirements (e.g., consistency, accuracy). In medical research, methodological criteria measure evidence quality, thus assessing a study's validity. Many studies about medical document classification abide by evidence quality criteria. To assess the efficacy of a treatment drug, study subjects could be randomly divided into experiment and control groups. If patient assignment is not properly based on sound principles, i.e., people in the experiment group are healthier and younger than those in the control group, we cannot trust the conclusions drawn from that study. Reliability is a matter of common interest in every domain, but evidence quality is especially important in medicine. Evidence-based medicine has thus been heavily emphasized.

Manual classification. Considering the vast number of articles published each year, it is difficult for clinicians to keep up with recent advances in medicine. Several meta-journals deliver high-quality research articles in medicine, including the *ACP Journal Club* [2], *Cochrane Reviews* [4], and *Evidence-Based Medicine* [5]. For these meta-journals, reviewers who are experts in specific areas regularly evaluate research articles, choosing the most significant studies on specific research topics and writing reviews with summaries or citations. Although the meta-journals regularly deliver high-quality information, manually searching for and reviewing articles requires the hard work of many experts.

Hand-crafted rules. In 1994, Haynes et al. developed a Boolean search filter for *PubMed* and updated it in 2005 with the *Clinical Hedges Database* (CHD) [34,35]. They evaluated candidate terms that could be used to retrieve high-quality articles in a Boolean search system for specific clinical tasks. CHD was used to develop the *PubMed Clinical Queries* [10], which were designed to act like 'hedges' against primary search results on *PubMed*.

Demner-Fushman et al. tried to extract clinically relevant information from MEDLINE abstracts for the clinical question answering task [23]. Starting from the initial *PubMed* search results, they tried to give relevant abstracts higher ranks and generate answers from these relevant abstracts. To measure relevance from the EBM perspective, the authors designed a direct scoring function: $S_{EBM} = S_{PICO} + S_{SoE} + S_{task}$; S_{PICO} represents contributions from matching PICO structures, e.g., Problem, Intervention, S_{SoE} represents evidence strength, and S_{task} represents factors associated with search task, e.g., treatment, diagnosis. They developed a series of knowledge extractors to automatically calculate each component: S_{SoE} is calculated as $S_{SoE} = S_{journal} + S_{study} + S_{date}$, $S_{journal}$ represents the impact of the published journal, S_{study} represents the study type, e.g., clinical trial, randomized controlled trial, and S_{date} represents the recency factor. Although each element reflects what physicians consider when they search for a MEDLINE citation, much time and domain knowledge is required to properly optimize system performance when combining those factors into a single ranking method. To evaluate the ranking results, the authors only assessed topical relevance, which differs from our test collection (Section 3.2).

Machine learning methods. Aphinyanaphongs et al. applied machine learning methods to identify high-quality and content-specific articles [16]. They assume that any article that is abstracted or cited in the *ACP Journal Club* is a high-quality article on a specific subject. The method of selecting methodologically rigorous articles adopted by the *ACP Journal Club* is the same as that adopted by CHD [1,34]. A machine-learning classifier used the document's title, abstract, MeSH, and publication type fields. The authors built a machine-learning classifier using a support vector machine (SVM), Naïve Bayes and text-specific boosting. Their model showed better or comparable performance to Boolean methods (*PubMed Clinical Queries*).

Kilicoglu et al. also used machine learning techniques to classify scientifically rigorous articles in CHD [50]. They experimented with various feature-classifier combinations. They exploited high-level semantic features, utilizing a knowledge-based system, SemRep [64]. They experimented with Naïve Bayes, SVM, boosting, and stacking methods as classifiers. In their experiment, the stacking method, which combines individual feature classifiers, achieved better performance than individual classifiers. In our study, we did not use semantic tools or a sophisticated meta-classifier because our system showed satisfactory classifier performance on its cross-validation test (Section 3.3.2).

2.4. Beyond the topical relevance

A search engine aims to bring relevant documents to its users. A retrieval model is a formal representation of the process of matching a query and a document, and it is the basis of the ranking algorithm [20]. Relevance is, however, assessed relative to an information need, not a query [56]. A user's information need may sometimes be more comprehensive, so it lies beyond the notion of topical relevance.

One example is the TREC Blog track opinion finding task [42,55]. The opinion finding task is an articulation of a user search task, where the required information is an opinion or has a perspective-finding nature, instead of a fact-finding nature (Fig. 1).

```
<top>
<num> Number: 930 </num>
<title> ikea </title>
<desc> Description:
Find opinions on Ikea or its products.
</desc>
<narr> Narrative:
Recommendations to shop at Ikea are relevant opinions. Recommendations of Ikea
products are relevant opinions.Pictures on an Ikea-related site that are not
related to the store or its products are not relevant.
</narr>
</top>
```

Fig. 1. An example of a TREC Blog track 2007 opinion retrieval task [55].

Most participants in the Blog track used a two-stage approach [29,55,77,83,84]. In the first stage, they performed relevance ranking using various weighting models, including BM25, Divergence from Randomness, or Language Modeling. In the second stage, the retrieved documents were reranked counting in opinion finding features.

A previous study [18] defined the metasearch problem as combining ranked lists of documents returned by multiple search engines in response to a given query to optimize combination performance. In this paper, we applied fusion techniques (e.g., Borda-fuse) to effectively combine different aspects of a user's information need into a single ranking process. In medical research, the importance of evidence quality is widely accepted, so we can tacitly assume that users care about a document's evidence quality as much as its topical relevance.

3. Method

3.1. Overall ranking strategy

We designed a ranking strategy as a three-step process (Fig. 2). First, we measured the relevance score for each document using a probabilistic retrieval model (Okapi BM25). Second, we used a machine-learning classifier to compute the quality score. We experimented with Naive Bayes, SVM^{light}, and SVM^{perf} as machine-learning classifiers. Finally, we combined the relevance and quality scores, using various fusion methods to draw the final ranking scores.

3.2. Test collections

We used two different text collections in our experiments (Fig. 3). The main features and preparations of each test collection are described below.

3.2.1. EBM test collection

Previous research used *OHSUMED* to evaluate query relevance in biomedical research and CHD to evaluate literature classifiers based on quality criteria [40]. To the best of our knowledge, no suitable collection was available for comprehensive evaluation that meets both criteria. To evaluate the final search results for both query-relevance and methodological quality, we built a test collection (*EBM test collection*) that utilizes preexisting sources.

The 2009 MEDLINE/PubMed citation database was used as a corpus. The *Cochrane databases of systematic reviews* (*Cochrane Reviews*) were utilized for hypothetical, but very reasonable, queries and gold standards. Table 1 and Fig. 4 describe the *EBM test collection* specifications.

3.2.1.1. Corpus. We leased *2009 MEDLINE®/PubMed®Journal Citations* from the U.S. National Library of Medicine [7]. The journal citations created in December 2008 contain 17,764,826 MEDLINE citations. Because these citations comprise most of the MEDLINE citations used by *PubMed*, our corpus is close to the real environment in its search scope.

3.2.1.2. Cochrane Reviews. Cochrane Reviews [4] are systematic reviews on research in human health care and policy. They collect rigorous and up-to-date research. The *Cochrane Reviews* primarily focus on the effectiveness of healthcare treatments

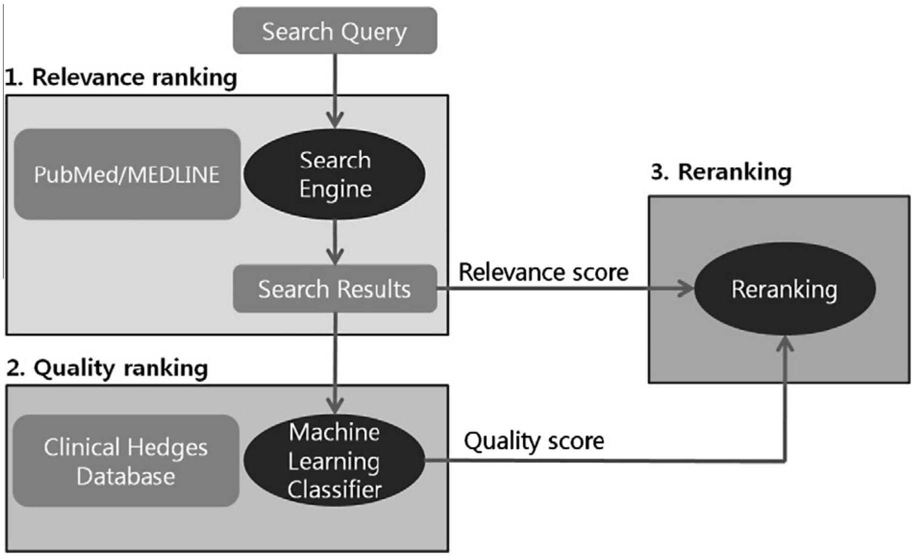


Fig. 2. Ranking strategy in this study.

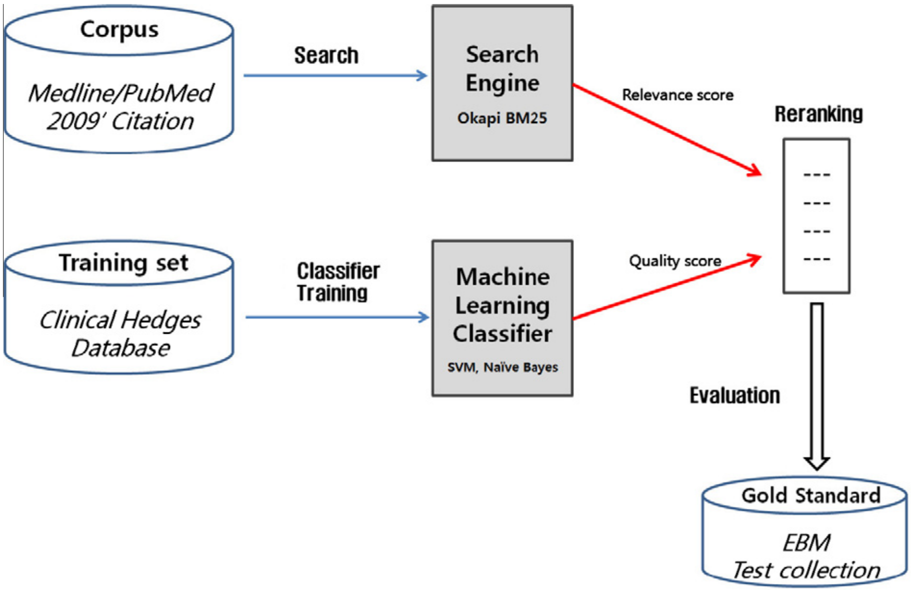


Fig. 3. Utilizing test collections for this study: The *EBM test collection* was used as a gold standard for evaluation. *Medline/PubMed 2009' Citation* was adopted as the corpus for this test collection. The *Clinical Hedges Database* was used only for quality ranking.

Table 1
Specifications of our test collection.

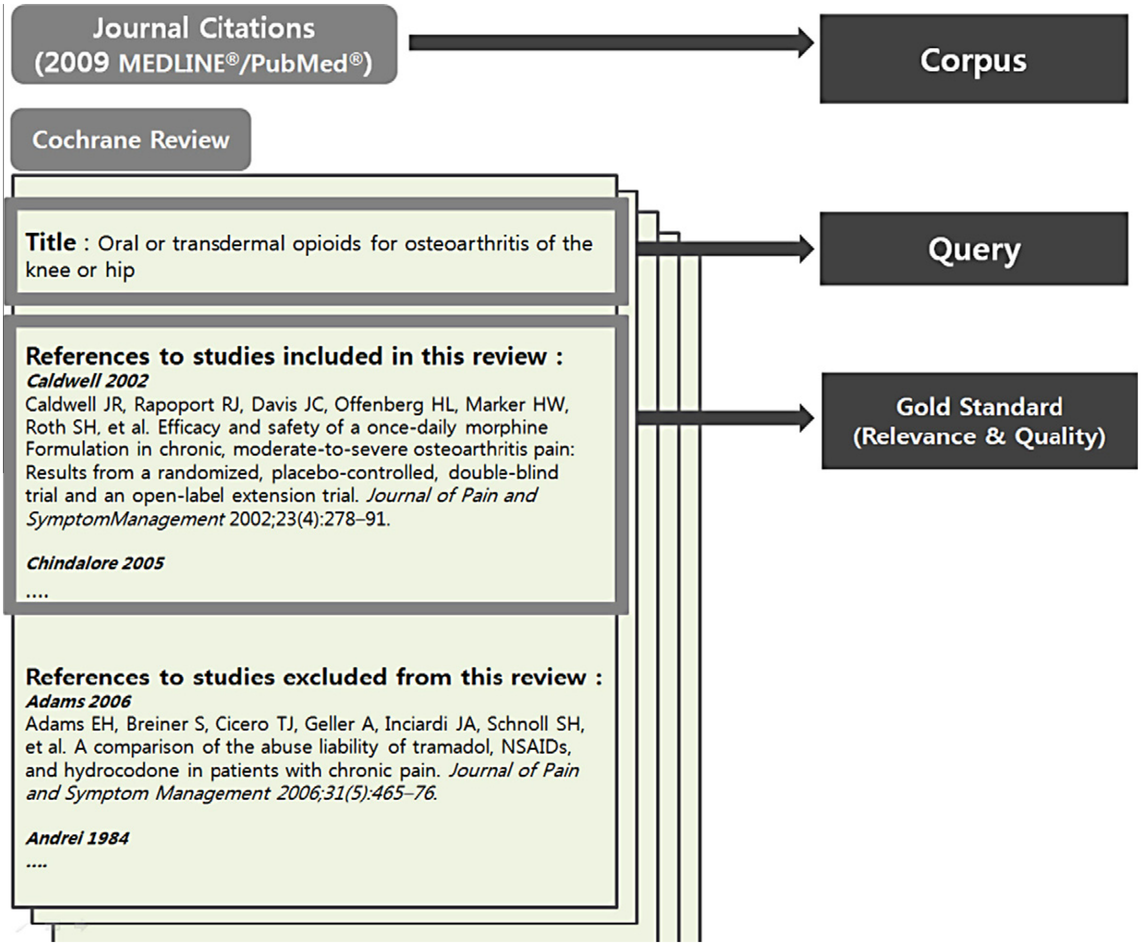
Source	Test collection	Number
Journal Citations (*2009 MEDLINE®/PubMed®*)	Corpus	17,764,826
Title of *Cochrane Reviews*' article	Queries	145
References Included, *References Excluded*	Documents Satisfying *Relevance* criteria	4,230 query-document pair
References Included	Gold Standard (Documents Satisfying both *Relevance and Quality* criteria)	1,559 query-document pair

and interventions as well as methodology and diagnostic tests. Each *Cochrane Review* is assigned to one of three types: intervention reviews assess the benefits and potential harm of interventions, methodology reviews address issues pertaining to how systematic reviews and clinical trials are conducted and reported, and diagnostic test accuracy reviews assess how well a diagnostic test performs in diagnosing and detecting a specific disease [3]. In this experiment, we have chosen only intervention reviews for the *EBM test collection*.

In each review, authors explicitly describe their objectives, the search strategies used for each database, their selection criteria, their data collection and analysis methods, and the results, discussion, conclusions, and references. They perform a comprehensive search for all potentially relevant studies for a given topic. After comprehensively collecting the relevant studies on this topic, the authors draw conclusions based on the trial results that meet predefined quality criteria. From the retrieved articles, references that meet the selection criteria are listed in the '*References Included*' category, with the characteristics of the study described (Table 2). References that do not meet the eligibility criteria are listed in the '*References Excluded*' category, with the reasons for the exclusion described (Table 3).

We took each *Cochrane Reviews* article's objective (e.g., ''To examine the effects of doxycycline compared with placebo or no intervention on pain and function in patients with osteoarthritis of the hip or knee.'') and hypothesized it as a possible user search intent (this experiment did not directly use these objectives). The topics in a *Cochrane Reviews* article (e.g., ''Doxycycline for osteoarthritis of the knee or hip'') were adopted as search queries. The topics are implicative in nature and have fewer words (9.3 words on average in the *EBM test collection*) than the objectives. They are thus more appropriate for use as queries.

In medical information retrieval, PICO (Patient problem, Intervention, Comparison, and Outcome) elements are regarded as major concepts in query formulation [73]. These elements are sometimes extended to PICOTT, adding Type of question and Type of study design information [68]. If we examine the search strategies used in a *Cochrane Reviews* article (Table 4), authors of the article formulated these concepts into a complex Boolean search strategy, utilizing wild cards and Boolean operators. Building an effective Boolean query strategy takes lots of time and experience. Conversely, topics in a *Cochrane Reviews* article generally contain only patient problem and intervention aspects, which is the core information compared to other elements, and the topics do not require much time for user to determine a query. This study aims for a more intel-



Title : Oral or transdermal opioids for osteoarthritis of the knee or hip

References to studies included in this review :

Caldwell 2002

Caldwell JR, Rapoport RJ, Davis JC, Offenberg HL, Marker HW, Roth SH, et al. Efficacy and safety of a once-daily morphine Formulation in chronic, moderate-to-severe osteoarthritis pain: Results from a randomized, placebo-controlled, double-blind trial and an open-label extension trial. *Journal of Pain and SymptomManagement* 2002;23(4):278–91.

Chindalore 2005

....

References to studies excluded from this review :

Adams 2006

Adams EH, Breiner S, Cicero TJ, Geller A, Inciardi JA, Schnoll SH, et al. A comparison of the abuse liability of tramadol, NSAIDs, and hydrocodone in patients with chronic pain. *Journal of Pain and Symptom Management 2006;31(5):465–76.*

Andrei 1984

....

Fig. 4. Test collection components.

ligent search system that reduces the user's burden, so using topics in *Cochrane Reviews* as queries is a natural, choice for our target use case.

We assumed that both *References Included* and *References Excluded* were relevant to the search intent but that only the *References Included* fulfilled the methodological quality criteria. A comprehensive search and the rigorous assessment efforts made by expert reviewers identified articles in *References Included*. We hypothesized that we could confidently say that we rank the documents in the same way as EBM-principled expert reviewers if we could automatically rank high-quality articles higher in the search results. Among the *Cochrane Reviews* published in 2009 and 2010, we arbitrarily picked 153 articles that were categorized as intervention reviews and most recently updated in 2009. *Cochrane Reviews* are updated regularly to reflect new evidence, as the results in new studies can change the conclusions of a review [3]. Because our corpus has MEDLINE citations indexed until December 2008, we checked the *Cochrane Reviews*' '*Last assessed as up-to-date*' field and selected articles that were last updated between January 2009 and December 2009. We could thus be certain that the 17 million MEDLINE citations were all made available to the expert reviewer before the last update in the *Cochrane Reviews*.

Each *MEDLINE/PubMed* citation record has a *PubMed Unique Identifier* (PMID). We hired 2 undergraduate students and trained them to search for *References Included* and *References Excluded* articles on *PubMed* (5,582 total articles). Every not-found article was checked twice by the first author. Because *Cochrane Reviews*' search range is not limited to *PubMed/MEDLINE*, there were references not indexed in *PubMed*. In PMID, 4,457 (80%) articles were found. After removing duplicates, we had 4,230 total query-PMID pairs.

Among the 153 *Cochrane Reviews*, 8 reviews did not have any *References Included* or *References Excluded*. Excluding these 8 reviews, we had 145 reviews that functioned as queries and answers. We randomly assigned 100 queries to the training set, and the remaining 45 queries were assigned to a held-out test set. The mean average precision (MAP) was chosen as our evaluation metric. Only *References Included* (1,559 query-PMID pairs found, which is 10.8 documents per query) was assumed to be our gold standard (Table 1).

Table 2
One example of the 'Characteristics of included studies' section of the *Cochrane Review 'Oral or transdermal opioids for osteoarthritis of the knee or hip'* [61].

Caldwell 2002	
Methods	Randomized controlled trial 4-arm parallel group design Trial duration: 4 weeks Multicentre trial No power calculation reported
Participants	Patients with prior suboptimal analgesic response to NSAIDs/paracetamol or previous intermittent opioid therapy were eligible; 295 patients with knee and/or hip osteoarthritis were reported at baseline; Number of females:184 of 295 (62%); Average age: 62 years
Intervention	Experimental interventions (a) oral morphine (Avinza), 30 mg once daily in the morning (b) oral morphine (Avinza), 30 mg once daily in the evening (c) oral morphine sulphate (Contin), 15 mg twice daily Control Intervention Placebo, twice daily Treatment duration: 4 weeks No analgesics other than study drugs allowed
Outcomes	Extracted pain outcome: global pain after 4 weeks Extracted function outcome: WOMAC disability subscore after 4 weeks Primary outcome: WOMAC OA index

Item	Authors' judgement	Description
Risk of Bias		
Adequate sequence generation?	Unclear	No information provided
Allocation concealment?	Unclear	No information provided
Described as double-blind?	Yes	
Blinding of patients?	Yes	
Blinding of physicians?	Unclear	No information provided
Blinding of outcome assessors?	Yes	
Interventions reported as indistinguishable?	No	
Double-dummy technique used?	Yes	
Intention-to-treat analysis performed? Pain	No	No information on exclusions available
Intention-to-treat analysis performed? Function	No	No information on exclusions available
No funding by commercial organization?	No	Sponsor: Elan

Table 3
Example from the 'Characteristics of excluded studies' from *Cochrane Reviews* 'Oral or transdermal opioids for osteoarthritis of the knee or hip'.

Adams 2006	Only active control interventions
Andrei 1984	Percentage of patients with knee or hip osteoarthritis 17% (5/30)
Burch 2004	No randomised controlled trial
Brooks 1982	Percentage of patients with osteoarthritis 50%, no information about joints involved
	...omitted for space reasons...
Wallace 1994	Crossover trial providing pooled results only
Wang 1965	Percentage of patients with osteoarthritis 6% (2/34)

3.2.2. Clinical Hedge Database

CHD is made for developing effective search strategies in *PubMed*. Using it, researchers can evaluate strategies for retrieving studies that are methodologically rigorous, reliable, and clinically relevant from biomedical research bibliographic databases. CHD contains 49,028 articles from 161 clinical journals that were published in 2000. Experts on the Hedges Team manually reviewed and classified documents according to the following aspects: format categories (e.g., original study, review article, general article, case report), purpose categories (e.g., causation, prognosis, diagnosis, therapy, costs, economics, clinical prediction guide, qualitative, something else), and methodological rigor (e.g., methodological rigor criteria for therapy, random allocation of participants to comparison groups, outcome assessment of at least 80% of those entering the investigation in one major analysis at any given follow-up assessment, and analysis consistent with study design).

In this study, we used CHD as the document quality classifier's training set. Documents that satisfied the following three constraints were taken as target documents: (1) categorized as an original study or review article, (2) belongs to the treatment task (in the *EBM test collection*, all queries pertain to the treatment task), and (3) fulfills the methodological rigor criteria. There were 2228 (5%) target documents in the CHD.

Table 4
Sample search strategy from *Cochrane Reviews* 'Doxycycline for osteoarthritis of the knee or hip' [60].

Search terms for design
1. randomized controlled trial.pt
2. controlled clinical trial.pt
3. randomized controlled trial.sh
...omitted for space reasons...
Search terms for Osteoarthritis
20. exp osteoarthritis/
21. osteoarthriti$.ti,ab,sh
22. osteoarthro$.ti,ab,sh
...omitted for space reasons...
Search terms for Doxycycline
31. exp doxycycline/
32. doxycycline.tw
33. deoxyoxytetracycline.tw
...omitted for space reasons...
Combining terms
45. or/1–19
46. or/20–30
47. or/31–44
...omitted for space reasons...
53. remove duplicates from 52

3.3. Preparing ranking modules at each step using the training set

3.3.1. Relevance ranking

We used the Okapi BM25 [65] probabilistic retrieval model implemented in *Terrier* [43]. We chose Okapi BM25 because it is a representative retrieval formula, and previous studies indicate that it shows more stable performance than other models, including the vector space or language models [27,67]. Accordingly, Okapi BM25 has been adopted as a baseline method in various information retrieval studies [32,55,66,74].

In the Okapi BM25 formula, top-ranked documents are retrieved by computing a similarity measurement between query q and document d [71]:

$$\sum_{t\in Q,D} \in \frac{N-df+0.5}{df+0.5} \times \frac{(k1+1)tf}{(k1(1-b)+b\frac{dl}{avdl})+tf} \times \frac{(k3+1)qtf}{k3+qtf},$$

where t is a term of query q, tf is the term's frequency in document d, qtf is the term's frequency in the query, N is the total number of documents in the collection, df is the number of documents contain the term t, dl is the document length, and $avdl$ is the average document length. Variables k_1, b, and k_3 are tuning parameters.

The title, abstract, Medical Subject Headings (MeSH) [8], and publication type fields were extracted from each *PubMed/Medline* citation and indexed, as shown in Fig. 5. We retrieved 1000 documents per query, with a 100,000 total documents in the training set. About half of the target documents (537 among 1198 query-document pair) were retrieved. The macro-averaged precision was 0.6%, and the macro-averaged recall was 56.2%. This study used these initially retrieved documents repetitively in subsequent processes (quality ranking and reranking stages). Only the relative ranking of each document, which MAP measured, changed within the same initially retrieved document sets during each query. The MAP in relevance ranking stage was 4.5%.

3.3.2. Quality ranking

Machine-learning classifiers were used to assess the methodological quality of documents retrieved from relevance ranking. We used 49,028 CHD documents as the quality classifier's training set, and 100,000 documents from the initial search results were used as a validation set for quality ranking. Fig. 6 illustrates the overall scheme.

The area under the ROC curve (AUC) was adopted as an evaluation metric. AUC has been widely used to measure classifier performance. A graph of sensitivity against 1 – specificity is a receiver operating characteristic (ROC) curve [19]. Points on the ROC curve are depicted by varying decision threshold. AUC offers a single classifier performance measure, invariant to the decision criterion selected [15].

We experimented with three different classifiers: Naïve Bayes [33], SVM^{light} [45], and SVM^{perf} [48]. SVM^{perf} is an implementation of the support vector machine formulation for optimizing multivariate performance measures [46]. In this experiment, we specified the loss function option in SVM^{perf} as the AUC.

Datasets: We converted MEDLINE citations into machine learning datasets as follows. We extracted and combined the title, abstract, MeSH, and publication type fields. After tokenization, case folding, stemming and stop word removal, a feature

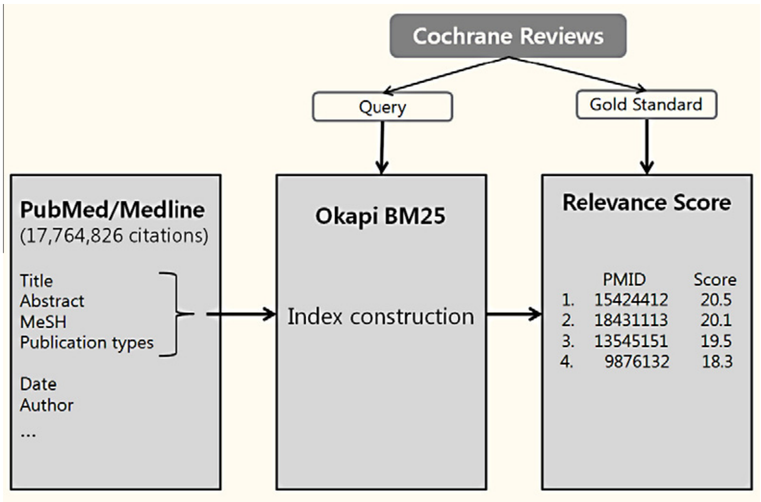


Fig. 5. Relevance ranking details.

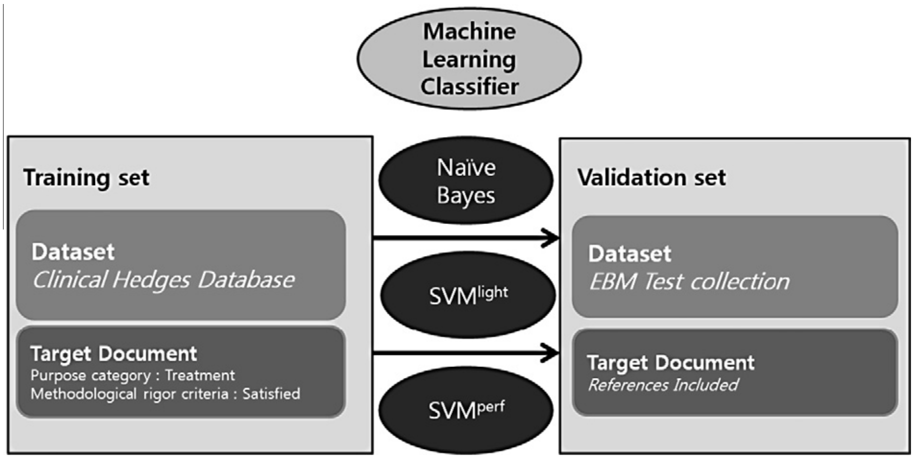


Fig. 6. General quality ranking scheme.

ID was assigned to each unique term, and the term frequency was used as a feature value. Each term was evaluated using the information gain method [79], and the top-scoring 1000 terms were chosen as a feature set.

Preliminary runs: We performed preliminary experiments within the training set to choose the number of features and the kernel function in SVM.

(1) Feature selection: We performed a 10-fold cross-validation with various numbers of features (1000, 2000, and 3000). The performance (AUC) gain was minimal (approximately 0.1%) when we increased the feature numbers to over 1000.
(2) Kernel function: We performed a 10-fold cross-validation with various kernel functions (linear, polynomial, sigmoid, rbf) and parameter combinations. The single best performance value for each kernel function was similar (AUC of approximately 98%). However, linear kernels showed more stable performance against altering parameter values than the other kernel functions.

Quality classifier training and validation: We generated various model sets using different classifier and parameter combinations (Table 5). Each model was trained on CHD. We attempted to find the best model, optimized for AUC, on the validation set. Because the quality score was designed as a feature value in the re-ranking process, the classification accuracy from the fixed threshold value did not matter. Our intent was to secure sufficient discriminatory power for a quality score by making the target documents attain a higher score than the non-target documents.

The SVM^{perf} (linear kernel, $c = 0.01$) classifier scored the best (AUC 89.6%) on the validation set (Fig. 7). As described in [46], we believe that this occurred because SVM^{perf}implements the alternative structural formulation of the SVM optimiza-

Table 5
AUC results on the validation set.

	Parameter value	AUC (%)
Naïve Bayes	*	79.7
SVM^{light} (linear kernel)	c = 0.0001	61.7
	c = 0.001	73.3
	c = 0.01	75.7
	c = 0.1	76.1
	c = 1	74.5
SVM^{perf} (linear kernel)	c = 0.0001	87.9
	c = 0.001	89.4
	c = 0.01	89.6
	c = 0.1	87.9
	c = 1	87.4

*: No parameter optimization is necessary for Naive Bayes; c: normalization constant.

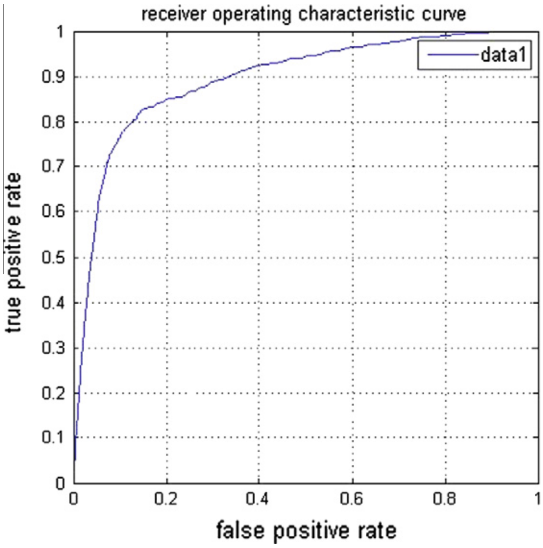


Fig. 7. SVM^{perf} classifier scored AUC 89.6% on the validation set.

tion problem, thus directly optimizing the ROC area. The decision function value for each document was used as a quality score.

3.3.3. Re-ranking

After computing the relevance and quality scores, we combined the two scores with various reranking methodologies. We used several simple combination methods [18,25,58] and SVM^{rank} [48], which used SVM algorithms to predict the rankings. Table 6 summarizes the re-ranked results on the training set. Our reranking methodologies are applied on identical lists (initially retrieving 1000 search results per query), so precision and recall rate have the same values. Only MAP varies with the reranking method.

Table 6
Re-ranked results on the training set.

	MAP (%)
Relevance ranking	4.5
Quality ranking	7.5
Re-ranking method	
Linear combination	7.8
Multiplicative combination	13.6
Borda-fuse	12.2
Weighted linear combination (1:5)	11.9
Weighted multiplicative combination (1:0.5)	14.2
Weighted Borda fuse (1:5)	14.0
SVM$^{rank(linear\ kernel,c=0.001)}$	12.1

Linear combination. Documents are re-ranked according to the sum of the relevance and quality scores.

$$\mathit{Reranking\ Score} = \mathit{Relevance\ score} + \mathit{Quality\ score} \tag{1}$$

Multiplicative combination. The relevance score is multiplied by the normalized quality score and re-ranked.

$$\mathit{Normalized\ Quality\ Score} = (\mathit{Quality\ score} - \mathit{Min})/(\mathit{Max} - \mathit{Min}) \tag{2}$$

$$\mathit{Reranking\ Score} = \mathit{Relevance\ score} * \mathit{Normalized\ Quality\ score} \tag{3}$$

Borda-fuse. Borda-fuse [18] is based on election strategies and requires neither the relevance score nor the training set. We compute the Borda-fuse score thus:

$$\mathit{Reranking\ Score} = 1/(\mathit{Relevance\ rank} + \mathit{Quality\ rank}) \tag{4}$$

Weighted linear combination. This combination is similar to the linear combination. The difference is that we can use the training data to optimize the weight factors with MAP.

$$\mathit{Reranking\ Score} = \alpha * \mathit{Relevance\ score} + \beta * \mathit{Quality\ score} \tag{5}$$

Among various (α: β) weight factors, (1:5) provided the best MAP (11.9%).

Weighted multiplicative combination. Among the various (α: β) weight factors, (1:0.5) gave the best MAP (14.2%).

$$\mathit{Reranking\ Score} = (\mathit{Relevance\ score})^{\alpha} * (\mathit{normalized\ Quality\ score})^{\beta} \tag{6}$$

Weighted Borda-fuse. Among various (α: β) weight factors, (1:5) presented the best MAP (14.0%).

$$\mathit{Reranking\ Score} = 1/(\alpha * \mathit{Relevance\ rank} + \beta * \mathit{Quality\ rank}) \tag{7}$$

SVM$^{\text{rank}}$. *Learning to rank* [53] refers to new ranking methodologies that leverage machine-learning technologies in the ranking process. The ranking model seeks to rank documents similar to the rankings in the training data. Many *Learning to rank* algorithms have been published. In this experiment, we used the SVM$^{\text{rank}}$ implementation.

SVM$^{\text{rank}}$ is an instance of SVM$^{\text{struct}}$ [75], which is used to efficiently train Ranking SVMs. In this experiment, SVM$^{\text{rank}}$ takes each document's relevance and quality scores as input features and predicts a final ranking as output. In the training set, a preference score for *References Included* was set to 1, and other documents were set to 0.

We performed a 10-fold cross-validation on the training set with various c parameters. We experimented with the linear kernel. We spent more than two weeks training SVM$^{\text{rank}}$ with other non-linear kernels, but the training job kept running without completion. SVM$^{\text{rank}}$ (linear kernel, $c = 0.001$) scored the best (MAP 12.6%) on the cross-validation test. When trained on the whole training set and reapplied, the MAP was 12.1%.

4. Results

We applied the aforementioned methodology to the held-out test set. At the relevance ranking stage, we obtained relevance-ranked retrieval results using the Okapi BM25 weighting model implemented in the *Terrier* search engine. The MAP was measured as 7.4%. The macro-averaged precision was 0.4%, and the macro-averaged recall was 56.0%. At the quality ranking stage, the SVM$^{\text{perf}}$ classifier trained on CHD was applied to the relevance ranked results, printing the quality score for each document. The AUC was measured at 93.5%.

Finally, we combined the relevance and quality ranking lists with various fusion methodologies to obtain the final ranking results.

We performed paired *t*-tests to verify the statistical significance of the MAP improvements from the baseline. Given two paired sets of measured values, the paired *t*-test determines whether they significantly differ from each other under the

Table 7
Re-ranked results on the test set.

	MAP (%)
Relevance ranking	7.40
Quality ranking *(Baseline)*	8.20
Re-ranking method	
Linear combination	13.0
Multiplicative combination	16.4[a]
Borda-fuse	19.6[a]
Weighted linear combination (1:5)	14.7[b]
Weighted multiplicative combination (1:0.5)	16.0[a]
Weighted Borda fuse (1:5)	16.0[a]
SVM$^{\text{rank(linear kernel,c=0.001)}}$	14.5[b]

[a] *p*-value less than 0.01.

[b] *p*-value less than 0.05.

assumptions that the paired differences are independent and identically normally distributed [30]. To compare MAP values between two different methods, the paired *t*-test is the preferred method due to its validity and power [22].

Table 7 summarizes the evaluation results. Borda-fuse scored best on a held-out test set. The Borda-fuse, weighted-Borda-fuse, multiplicative combination, and weighted multiplicative combination showed a significant increase in MAP (p-value < 0.01) compared to the baseline. The weighted linear combination and SVM^{rank} also showed some improvement (p-value < 0.05). Overall, there was a twofold increase in MAP values after applying the fusion technique.

5. Discussion

Ranking and Re-ranking. Compared to relevance or quality ranking alone, our re-ranking methodologies increased the performance impressively, showing great potential. We can summarize our future directions as follows.

5.1. Integrating other useful features

In this study, we combined two different features: Okapi BM25 relevance score and quality classifier score.

For the relevance aspect, we retrospectively tried other retrieval models with *Terrier* implements (TF-IDF, DFR_BM25 [13]) on the training set, but there was no significant difference in MAP (both measured as 4.6%) compared to the Okapi BM25 model (4.5%). Okapi BM25 is generally regarded as the most effective retrieval function, performing similar to the pivoted normalization vector space [72] and Dirichlet prior language models [82], which generally have similar performance when properly optimized [81]. Okapi BM25 and the other models above are based on the *bag-of-words* assumption; we believe that it is time to look beyond the *bag-of-words* model to bring significant advancement in ranking. Various studies [28,37,44,51,54,62] have attempted to utilize syntactic or semantic information in a retrieval model, which *bag-of-words* models ignored. Sensibly designing and incorporating these features in our re-ranking framework will be the focus of our next study.

5.2. Securing a certain recall level

The initial relevance ranking on the training set scored 4.5% on MAP and 7.4% on the held-out test set. We observed that roughly one-fourth of the gold standard documents ranked within the top 100, and another one-fourth was located between the top 101 and 1000. The remaining half was outside of the top 1000, excluding themselves from our reranking processes. Achieving a certain recall rate is important for medical researchers or policy makers who give higher priorities to search comprehensiveness over convenience. This problem is difficult, considering the denominator size. As a first step, we consider utilizing a Boolean model with a ranked retrieval model [49].

5.3. Comparing various re-ranking methods

In the TREC Blog track, a previous study [29] performed experiments using SVM^{map} [80] with two different features (relevance and opinion scores) and default parameter settings. Rank Learning outperformed other simple combination methods, including Borda-fuse. In this study, simple combination methods showed better or comparable results to SVM^{rank}, which somewhat differs from the previous study [29], suggesting that the optimal re-ranking methodology might differ depending on the field to which it is applied or the feature characteristics. We plan to perform expanded-scale experimentation when the test collection enlargement is complete to draw more solid conclusions between various fusion methodologies.

5.4. Quality classifier performance generality

In prior studies regarding high-quality article classification [16,34–36,50], researchers evaluated classifiers within the same collection (using either a cross-validation or a held-out test set). In this study, we used two separate collections; CHD was used as a training set, and the *EBM test collection* was made based on *Cochrane Reviews*. Although CHD and *Cochrane Reviews* share the EBM perspective, there are some differences between them. In *Cochrane Reviews*, the selection criteria are more specific, depending on the subject. Furthermore, the articles have different reviewers for each subject. However, the quality classifier (SVM^{perf}, linear kernel, $c = 0.01$, trained on CHD) showed relatively good performance on the held-out test set (AUC 93.5%). We are curious about the robustness or generality of the quality-ranking classifier when trained and applied on different collections. We are building another test collection for quality classification utilizing the *ACP Journal Club*, similar to Aphinyanaphongs et al. [16]. With three different test collections (CHD, *Cochrane Reviews*, *ACP Journal Club*) available, future experiments might provide further insights.

Test collection. We utilized *Cochrane Reviews* as our test collection. Using reliable preexisting sources, we can avoid the credibility issue and the heavy burden of manual review. We can summarize the strengths of our test collection in two different ways. First, 17 million MEDLINE citations were adopted as a corpus. This collection contains all *PubMed/MEDLINE* citations until December 2008. Second, having both query relevance and the quality aspects of a gold standard made robust evaluation possible.

Currently, we are augmenting the test collection with more *Cochrane Reviews* articles (i.e., adding more queries to the test collection). We hope to make this test collection public in the near future.

Every approach has both merits and demerits. This collection also has several limitations: (1) single-purpose category, i.e., most *Cochrane Reviews* belong to the intervention review category, so our test collection is confined to the treatment purpose category; and (2) lack of citation information, i.e., an important factor that *Google Scholar* uses in their ranking algorithm is 'what other articles have said about it'. Because full-text articles are not available for us, citation information is lost in our test collection.

6. Conclusions

In this paper, we attempted to design an effective EBM ranking algorithm. We combined relevance and quality ranking using various fusion methods, yielding significant improvements in the final ranking performance. In our study, we built our test collection utilizing *Cochrane Reviews*, using 17 million MEDLINE documents as a corpus, which met both relevancy and quality standards.

We are indebted to the prior studies that gave helpful insights, but we can derive inspiration from more studies. To evaluate study quality, we could not utilize powerful methods based on citation information, including h-index [11] and hg-index [12], due to the lack of citation information in our test collection. When we incorporate this aspect into our future test collection, more effective evaluation will be attainable. Other domains that consider the information overload problem include attempts to build recommender systems [52,63,70] by combining users' personalization characteristics. Referring to these approaches, we hope to make our ranking system more intelligent when we extend our research in our future studies.

Acknowledgments

Use of the Clinical Hedges database was made possible through a collaboration agreement with R.B. Haynes and N.L. Wilczynski at McMaster University, Hamilton, Ontario Canada. This work was supported by the National Research Foundation of Korea (NRF) with a grant funded by the Korean Government (MEST) (No. 2009-0075089).

References

[1] About ACP Journal Club (cited 2011 July 15), <http://www.acpjc.org/shared/purpose_and_procedure.htm>.
[2] ACP Journal Club (cited 2011 July 15), <http://acpjc.acponline.org/>.
[3] The Cochrane Library (cited 2011 July 15), <http://www.thecochranelibrary.com>.
[4] Cochrane Reviews (cited 2011 July 15), <http://www.cochrane.org/cochrane-reviews>.
[5] Evidence-Based Medicine (cited 2011 July 15), <http://ebm.bmj.com/>.
[6] Google Scholar (cited 2011 July 14), <http://scholar.google.com>.
[7] Leasing Journal Citations (cited 2011 July 15), <http://www.nlm.nih.gov/databases/journal.html>.
[8] MeSH (cited 2011 July 15), <http://www.ncbi.nlm.nih.gov/mesh>.
[9] PubMed (cited 2011 July 14), <http://www.ncbi.nlm.nih.gov/pubmed/>.
[10] PubMed Clinical Queries (cited 2011 July 15), <http://www.ncbi.nlm.nih.gov/pubmed/clinical>.
[11] S. Alonso et al, h-Index: a review focused in its variants, computation and standardization for different scientific fields, Journal of Informetrics 3 (4) (2009) 273–289.
[12] S. Alonso et al, hg-Index: a new index to characterize the scientific output of researchers based on the h-and g-indices, Scientometrics 82 (2) (2010) 391–400.
[13] G. Amati, C.J. Van Rijsbergen, Probabilistic models of information retrieval based on measuring the divergence from randomness, ACM Transactions on Information Systems (TOIS) 20 (4) (2002) 357–389.
[14] Amit Ghosh, D. Stengel, Nancy Spector, Narayana Murali, Franz Porzsolt (Eds.), Evidence-Based Health Care Seen From Four Points of View, Optimizing Health: Improving the Value of Healthcare Delivery, 2006, pp. 205–216.
[15] P.B. Andrew, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition 30 (7) (1997) 1145–1159.
[16] Y. Aphinyanaphongs et al, Text categorization models for high-quality article retrieval in internal medicine, Journal of the American Medical Informatics Association 12 (2) (2005) 207–216.
[17] C. Apté, F. Damerau, S.M. Weiss, Automated learning of decision rules for text categorization, ACM Transactions on Information Systems (TOIS) 12 (3) (1994) 233–251.
[18] J.A. Aslam, M. Montague, Models for metasearch, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New Orleans, Louisiana, United States, 2001, pp. 276–284.
[19] V. Bewick, L. Cheek, J. Ball, Statistics review 13: receiver operating characteristic curves, Crit Care 8 (6) (2004) 508–512.
[20] D.M. Bruce Croft, Trevor Strohman (Eds.), Search Engines: Information Retrieval in Practice, Addison-Wesley Publishing Company, 2009, pp. 235–236.
[21] A.L. Cochrane (Ed.), Effectiveness and Efficiency: Random Reflections on Health Services, Nuffield Provincial Hospitals Trust, London, 1972, pp. 1–60.
[22] G.V. Cormack, T.R. Lynam, Validity and power of *t*-test for comparing MAP and GMAP, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Amsterdam, The Netherlands, 2007, pp. 753–754.
[23] D. Demner-Fushman, J. Lin, Answering clinical questions with knowledge-based and statistical techniques, Computational Linguistics 33 (1) (2007) 63–103.
[24] H. Drucker, D. Wu, V.N. Vapnik, Support vector machines for spam categorization, IEEE Transactions on Neural Networks 10 (5) (1999) 1048–1054.
[25] A. Edward, J.A.S. Fox, Combination of multiple searches, in: Proceedings of the 2nd Text REtrieval Conference (TREC-2), NIST Special Publication Gaithersburg, Maryland, USA, 1994.
[26] S. Eyheramendy, D.D. Lewis, D. Madigan, On the naive bayes model for text categorization, in: 9th International Workshop on Artificial Intelligence and Statistics, Key West, Florida, 2003, pp. 332–339.
[27] H. Fang, T. Tao, C. Zhai, A formal study of information retrieval heuristics, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Sheffield, United Kingdom, 2004, pp. 49–56.
[28] H. Fang, C.X. Zhai, Semantic term matching in axiomatic approaches to information retrieval, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2006, pp. 115–122.

[29] S. Gerani, M. Carman, F. Crestani, Investigating learning approaches for blog post opinion retrieval, in: M. Boughanem et al. (Eds.), Advances in Information Retrieval, Springer, Berlin/ Heidelberg, 2009, pp. 313–324.
[30] C.H. Goulden, Methods of Statistical Analysis, second ed., Wiley, New York, 1952.
[31] Y. Guo, Z. Shao, N. Hua, Automatic text categorization based on content analysis with cognitive situation models, Information Sciences 180 (5) (2010) 613–630.
[32] H. Zaragoza, N. Craswell, M. Taylor, S. Saria, S. Robertson, Microsoft cambridge at TREC 13: Web and hard tracks, in: Text Retrieval Conference, Gaithersburg, Maryland, 2004.
[33] M. Hall et al, The WEKA data mining software: an update, SIGKDD Exploration Newsletter 11 (1) (2009) 10–18.
[34] R.B. Haynes et al, Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey, BMJ 330 (7501) (2005) 1179.
[35] R.B. Haynes et al, Developing optimal search strategies for detecting clinically sound studies in MEDLINE, Journal of the American Medical Informatics Association 1 (6) (1994) 447–458.
[36] R.B. Haynes, N.L. Wilczynski, Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey, BMJ 328 (7447) (2004) 1040.
[37] A. Herdagdelen et al, Generalized syntactic and semantic models of query reformulation, in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Geneva, Switzerland, 2010.
[38] E. Herrera-Viedma et al, Evaluating the information quality of web sites: a methodology based on fuzzy computing with words, Journal of the American Society for Information Science and Technology 57 (4) (2006) 538–549.
[39] E. Herrera-Viedma, E. Peis, Evaluating the informative quality of documents in SGML format from judgements by means of fuzzy linguistic techniques based on computing with words, Information Processing & Management 39 (2) (2003) 233–249.
[40] W. Hersh et al, OHSUMED: an interactive retrieval evaluation and new large test collection for research, in: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Springer-Verlag New York, Inc., Dublin, Ireland, 1994, pp. 192–201.
[41] T. Hughes, An interview with Anurag Acharya, Google Scholar lead engineer (cited 2011 July 14), <http://www.aardvarknet.info/access/number59/monthnews.cfm?monthnews=04>.
[42] Iadh Ounis, Craig Macdonald, Ian Soboroff. On the TREC blog track, in: AAAI International Conference on Weblogs and Social Media, 2008.
[43] Iadh Ounis, G. Amati, Vassilis Plachouras, Ben He, Craig Macdonald, Christina Lioma, Terrier: a high performance and scalable information retrieval platform, in: ACM SIGIR Workshop on Open Source Information Retrieval, Seattle, WA, USA, 2006, pp. 18–25.
[44] N.C. Ide, R.F. Loane, D. Demner-Fushman, Essie: a concept-based search engine for structured biomedical text, Journal of the American Medical Informatics Association 14 (3) (2007) 253–263.
[45] T. Joachims, Making large-Scale SVM learning practical, in: Advances in Kernel Methods – Support Vector Learning, MIT Press, 1999, pp. 169–184.
[46] T. Joachims, SVM-perf, 2009 (cited 2011 July 15), <http://www.cs.cornell.edu/People/tj/svm_light/svm_perf.html>.
[47] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: European Conference on Machine Learning, Chemnitz, Germany, 1998, pp. 137–142.
[48] T. Joachims, Training linear SVMs in linear time, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Philadelphia, PA, USA, 2006, pp. 217–226.
[49] S. Karimi et al, The challenge of high recall in biomedical systematic search, in: Proceeding of the Third International Workshop on Data and Text Mining in Bioinformatics, ACM, Hong Kong, China, 2009, pp. 89–92.
[50] H. Kilicoglu et al, Towards automatic recognition of scientifically rigorous clinical research evidence, Journal of the American Medical Informatics Association 16 (1) (2009) 25.
[51] O. Kolomiyets, M.F. Moens, A survey on question answering technology from an information retrieval perspective, Information Sciences 181 (24) (2011) 5412–5434.
[52] D.R. Liu, P.Y. Tsai, P.H. Chiu, Personalized recommendation of popular blog articles for mobile applications, Information Sciences 181 (9) (2011) 1552–1572.
[53] T.-Y. Liu, Learning to rank for information retrieval, Foundations and Trends in Information Retrieval 3 (3) (2011) 225–331.
[54] Y. Lou, Z. Li, Q. Chen, Semantic relevance ranking for XML keyword search, Information Sciences 190 (0) (2011) 127–143.
[55] C. Macdonald, I. Ounis, I. Soboroff. Overview of the TREC 2007 blog track, in: Proc. of the Text Retrieval Conference, Gaithersburg, Maryland, USA, 2007.
[56] C.D. Manning, P. Raghavan, H. Schutze (Eds.), Introduction to Information Retrieval, vol. 1, Cambridge University Press, Cambridge, 2008, pp. 140–144.
[57] C.D. Manning, P. Raghavan, H. Schutze (Eds.), Introduction to Information Retrieval, vol. 1, Cambridge University Press, Cambridge, 2008, pp. 307–310.
[58] M. Montague, J.A. Aslam, Condorcet fusion for improved retrieval, in: Proceedings of the Eleventh International Conference on Information and Knowledge Management, ACM, McLean, Virginia, USA, 2002, pp. 538–548.
[59] T. Mullen, N. Collier, Sentiment analysis using support vector machines with diverse information sources, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Barcelona, Spain, 2004, pp. 412–418.
[60] E. Nuesch et al, Doxycycline for osteoarthritis of the knee or hip, Cochrane Database of Systematic Reviews 4 (2009).
[61] E. Nuesch et al, Oral or transdermal opioids for osteoarthritis of the knee or hip, Cochrane Database of Systematic Reviews 4 (2009) CD003115.
[62] J.H. Park, W.B. Croft, D.A. Smith, A quasi-synchronous dependence model for information retrieval, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, Glasgow, Scotland, UK, 2011, pp. 17–26.
[63] C. Porcel et al, A hybrid recommender system for the selective dissemination of research resources in a technology transfer office, Information Sciences 184 (1) (2012) 1–19.
[64] T.C. Rindflesch, M. Fiszman, The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text, Journal of Biomedical Informatics 36 (6) (2003) 462–477.
[65] S.E. Robertson, W.S. Okapi, Keenbow at TREC-8, in: Eighth Text REtrieval Conference (TREC-8), Gaithersburg, Maryland, 1999.
[66] M.J.T. Ruihua Song, Ji-Rong Wen, Hsiao-Wuen Hon, Yong Yu, Viewing term proximity from a different perspective, in: C. Macdonald et al. (Eds.), Advances in Information Retrieval, Springer, Berlin/ Heidelberg, 2008, pp. 346–357.
[67] J. Savoy, Data fusion for effective European monolingual information retrieval, in: C. Peters et al. (Eds.), Multilingual Information Access for Text, Speech and Images, Springer, Berlin/Heidelberg, 2005. 921-921.
[68] C. Schardt et al, Utilization of the PICO framework to improve searching PubMed for clinical questions, BMC Medical Informatics and Decision Making 7 (1) (2007) 16.
[69] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys (CSUR) 34 (1) (2002) 1–47.
[70] J. Serrano-Guerrero et al, A google wave-based fuzzy recommender system to disseminate information in University Digital Libraries 2.0, Information Sciences 181 (9) (2011) 1503–1516.
[71] A. Singhal, Modern information retrieval: a brief overview, IEEE Data Engineering Bulletin 24 (4) (2001) 35–43.
[72] A. Singhal, C. Buckley, M. Mitra, Pivoted document length normalization, in: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Zurich, Switzerland, 1996, pp. 21–29.
[73] R. Snowball, Using the clinical question to teach search strategy: fostering transferable conceptual skills in user education by active learning, Health Libraries Review 14 (3) (1997) 167–172.
[74] K.M. Svore, P.H. Kanani, N. Khan, How good is a span of terms? Exploiting proximity to improve Web retrieval, in: SIGIR, ACM, Geneva, Switzerland, 2010.

[75] I. Tsochantaridis et al, Support vector machine learning for interdependent and structured output spaces, in: Proceedings of the Twenty-first International Conference on Machine Learning, ACM, Banff, Alberta, Canada, 2004, p. 104.

[76] R. Vine, Google Scholar, Journal of the Medical Library Association 94 (1) (2006) 97–99.

[77] K. Yang, N. Yu, H. Zhang, WIDIT in TREC 2007 blog track: combining lexicon-based methods to detect opinionated blogs, in: Proceedings of the Text REtrieval Conference, Gaithersburg, Maryland, USA, 2007.

[78] Y. Yang, An evaluation of statistical approaches to text categorization, Information Retrieval, vol. 1(1), Springer, Netherlands, 1999, pp. 69–90.

[79] Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, in: Proceedings of the Fourteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., 1997, pp. 412–420.

[80] Y. Yue et al, A support vector method for optimizing average precision, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Amsterdam, The Netherlands, 2007, pp. 271–278.

[81] C. Zhai, A brief review of information retrieval models, Technical report, Department of Computer Science, University of Illinois at Urbana-Champaign, 2007.

[82] C. Zhai, J. Lafferty, A study of smoothing methods for language models applied to ad hoc information retrieval, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New Orleans, Louisiana, United States, 2001, pp. 334–342.

[83] Q. Zhang et al., Fdu at trec 2007: opinion retrieval of blog track, in: Proceedings of the Text REtrieval Conference, Gaithersburg, Maryland, USA, 2007.

[84] W. Zhang, C. Yu, W. Meng, Opinion retrieval from blogs, in: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, ACM, 2007, pp. 831–840.