Application of Information Technology ■

Evaluating Relevance Ranking Strategies for MEDLINE Retrieval

ZHIYONG LU, PHD, WON KIM, PHD, W. JOHN WILBUR, MD, PHD

Abstract This paper evaluates the retrieval effectiveness of relevance ranking strategies on a collection of 55 queries and about 160,000 MEDLINE® citations used in the 2006 and 2007 Text Retrieval Conference (TREC) Genomics Tracks. The authors study two relevance ranking strategies: term frequency–inverse document frequency (TF-IDF) weighting and sentence-level co-occurrence, and examine their ability to rank retrieved MEDLINE documents given user queries. Furthermore, the authors use the reverse chronological order—PubMed's default display option—as a baseline for comparison. Retrieval effectiveness is assessed using both mean average precision and mean rank precision. Experimental results show that retrievals based on the two strategies had improved performance over the baseline performance, and that TF-IDF weighting is more effective in retrieving relevant documents based on the comparison between the two strategies.

■ **J Am Med Inform Assoc.** 2009;16:32–36. DOI 10.1197/jamia.M2935.

Introduction

In this article, we investigate the capability of two relevance ranking strategies (term frequency–inverse document frequency [TF-IDF] vs. sentence-level co-occurrence) in the context of MEDLINE searches. Furthermore, the two strategies are compared to one another as well as to a baseline strategy.

Background

A number of techniques have been researched over the last 40 years in order to improve retrieval effectiveness including relevance ranking, query expansion, and relevance feedback in both the biomedical domain as well as other subject domains.[1–3] The work presented here examines one specific influential technique: relevance ranking as part of an ongoing investigation of how to improve upon MEDLINE search performance through PubMed.

Systems employing the relevance ranking technique attempt to sort retrieved documents based on some measures in order to display more relevant documents earlier. Two relevance ranking strategies investigated here are: TF-IDF weighting and sentence-level co-occurrence. They are compared to one another as well as to a baseline strategy that ranks retrievals by the reverse chronological order–the default display setting currently used in PubMed.

The TF-IDF ranking strategy is a simple *term weighting* approach in the vector space model where documents are represented as N-dimensional vectors (N is the number of unique terms in all documents) and are ranked according to the assigned weights of query terms. It was first developed by Salton[2] and has been extensively evaluated to show better retrieval output than the conventional Boolean approach normally used in operational retrieval situations.[3,4] In recent years, much IR research has focused on finding new and better term weighting schemes such as Okapi BM25.[5] Although some weighting schemes yielded better results in various tasks, none consistently outperformed any of the others over a range of different queries during a massive comparison of different weighting techniques against the TREC data.[6]

Sentence-level co-occurrence is a *term proximity* approach—a completely different heuristic for ranking documents—postulating that documents with query terms in close proximity are more relevant than documents with terms scattered throughout the document. The research on term proximity originated from early studies by Keen,[7,8] followed by some evaluation studies on TREC data sets.[9,10] Although improved performance was seen in some circumstances, results of these evaluation studies are not conclusive.[11]

Design Objectives

In addition to being used in the general information retrieval (IR) community, TF-IDF has also been applied to biomedical text retrieval recently. Lucene,[12] a search engine that implements TF-IDF weighting, was applied by multiple participating teams[13-15] in the previous TREC Genomics Tracks,[16,17] during which its ability to bring more relevant documents to the top of a ranked list has also been somewhat demonstrated. However, its precise contribution was difficult to characterize in those studies because a) it was often combined with other techniques in ad-hoc ways, and b) the lack of baseline performance. To this end, the first objective of this work is to evaluate the capability of the two ranking strategies in improving biomedical text retrieval while other techniques are kept constant or controlled for.

The second relevance ranking strategy, recently proposed by Siadaty et al.,[18] uses sentence-level co-occurrence as a surrogate for the existence of relationships between query words. More

Affiliation of the authors: National Center for Biotechnology Information (NCBI), National Library of Medicine, Bethesda, MD 20894.

Supported by the Intramural Research Program of NIH, National Library of Medicine. The authors are grateful to the TREC organizers for their efforts in producing and making the text collection and relevance judgments publicly available.

Correspondence: Zhiyong Lu, NCBI/NLM/NIH, 8600 Rockville Pike, Bethesda, MD 20852; e-mail: <luzh@ncbi.nlm.nih.gov>.

Received for review: 07/18/08; accepted for publication: 09/28/08.

specifically, the authors proposed that if words in a multi-word query occur within the same sentence in an article A versus different sentences in article B, then article A is more likely to be relevant to the original user query than B. Furthermore, they developed a search engine for MEDLINE called Relemed that employs such a strategy and showed that Relemed improved upon PubMed search performance.

However, their conclusion was drawn based on a very limited number of evaluation studies: two artificial queries in two case studies. As the authors pointed out, "additional evaluation and comparison of Relemed with PubMed and other search engines is essential" (for demonstrating its ability). This led to the second objective of our work, which is to evaluate this new ranking strategy by performing a large-scale comparison of this newly proposed strategy against the aforementioned TF-IDF weighting strategy as well as the baseline strategy. Note that instead of directly comparing results of some searches in PubMed and Relemed, we implemented our own search system that supports both Boolean and probabilistic text retrievals. Thus, not only can we evaluate the retrieval effectiveness of the strategies used in PubMed and Relemed, but we can compare them against the TF-IDF weighting strategy as well.

System Description

Text Collection

In this work, we first used the 2007 TREC Genomics data,[19] which consist of 36 topics (in the form of biological questions) and 162,048 full-text documents from Highwire Press (http://highwire.stanford.edu/). Except for 1,800 instances, most of the documents were successfully mapped to their corresponding PubMed Identifiers (PMIDs). Hereafter, we refer to the remaining set of 160,248 PMIDs as the *TREC set* in our study. The 36 topics (Topic IDs 200 to 235) were collected from bench biologists and represent actual information needs in biomedical research. For demonstration purposes, we show three topics 207, 229 and 231 in the list below:

<207> What toxicities are associated with etidronate?

<229> What signs or symptoms are caused by human parvovirus infection?

<231> What tumor types are found in zebrafish?

For each topic, a set of relevant documents from the TREC set were produced during the relevance judgments based on pooled results from team submissions. They are assumed as the ground truth or gold standard data in our investigation and referred to as the *relevant set* in the remainder of this paper. Since the same text collection was used in the 2006 TREC Genomics track,[20] the methods described here were also applied to the 28 topics in TREC 2006.

Automatic Query Construction

For each TREC topic, a user query is required for retrieving relevant documents in PubMed. In order to produce unbiased (i.e., without human interference) user queries in a consistent manner, we chose to automatically select words from questions as queries on the assumption that real users would also intuitively create their queries based on the questions. Specifically, for each question, we first removed stop words[21] from the question and enumerated all possible word combinations, each of which was then searched in PubMed. Subsequently, for each word combination that retrieved a non-empty set of documents, we compared those retrieved documents to the ones in the relevant set and computed the standard IR measures:[1] recall, precision and F-measure. In the end, the query with the highest F-measure was selected for studying the search strategies.

Take the topic 229 as example, we first removed stop words *what*, *or*, *are*, *by* from the question. A total of 63 different user queries were then generated based on the remaining six words (*signs*, *symptoms*, *caused*, *human*, *parvovirus*, *infection*) and subsequently searched in PubMed to obtain a set of relevant documents. Three example queries are shown in Table 1 together with their corresponding IR measures after comparing the retrieved set with the relevant set. The query *human parvovirus* was finally chosen to study the search strategies because it yielded the highest F-measure.

We processed all 36 topics in this way and were able to identify query terms for most of the topics except two instances (topics 207 and 225) where no query terms could be generated to represent meaningful user queries (i.e., their F-measures are almost zero). Therefore, the two topics were excluded from further analysis. The automatically generated queries for the remaining 34 TREC topics varied in length from a single word to a maximum of four words, with a mean of 2.4 words and median of 2 words per query.

Once user queries were determined, a list of documents was retrieved for each topic e.g., for topic 229, 42 PMIDs from the TREC set were returned using the query (*human parvovirus*) and these retrieved documents were then ranked by the three strategies below.

Ranking Strategies

Rank by Reverse Chronological Time Order

In order to rank documents by the reverse chronological order, we first sorted the PMIDs numerically and then reversed the order. As a result, larger PMIDs would appear earlier in a ranked list. Note that PMIDs are consistent with Entrez Dates (EDAT) but not with Publication Dates (PDAT). That is, a PMID reflects the time sequence of when a citation was first entered in PubMed, but not necessarily of when a citation was published. For instance, despite the fact that Wilbur et al., 2006[22] was published earlier than Lu et al., 2007,[23] the former (PMID: 18080004) is registered with a larger PMID than the latter (PMID: 17990498) due to its late entrance to PubMed (12/15/07 versus 11/10/07). Nevertheless, the result of our sorting mechanism is consistent with the default display order in PubMed (i.e., last in—first out). In summary, this kind of ordering chooses to display most recent citations earlier in order to prevent older citations from displaying near the top of retrievals.

Rank by TF-IDF Weighting

We computed TF-IDF weightings as follows based on our prior work.[24] The TF measure tf_{td} is assigned to each term t in each document d, in the following formula:

Table 1 ■ Three Potential User Queries Constructed Automatically by Selecting Words from Topic 229

Word Combination	Recall	Precision	F-measure
Human parvovirus	0.47	0.63	0.54
Human parvovirus infection	0.39	0.85	0.53
Parvovirus infection	0.39	0.69	0.49

Table 2 ■ The First Two Columns Show the Eight Relevance Levels Defined in Relemed. The Third Column Shows the Percentages and Counts of the Set of Retrieved Documents (3163 PMIDs) Associated with Each Level

Relevance Level	Query Must Match	Retrieved Documents
1	T and A and M	0.6% (18)
2	T and A	2.2% (69)
3	T and M	0.5% (16)
4	A and M	1.5% (48)
5	T	1.8% (57)
6	A	15.8% (501)
7	M	45.3% (1433)
8	T or A or M	32.3% (1021)
SUM		100% (3163)

T = title; A = at least one abstract sentence; M= concatenated MeSH terms.

$$tf_{td} = 1/\left(1 + \exp\left(\alpha \cdot d_{length}\right) \cdot \lambda^{f_{td}-1}\right) \quad (1)$$

where f_{td} denotes the frequency of term t within document d and d_{length} denotes the length of d. α and λ were previously determined as 0.0044 and 0.7, respectively.[25,26] The IDF measure IDF_t is assigned once for each term in the TREC set, in the following formula:

$$IDF_t = \log\left(N/n_t\right) \quad (2)$$

where n_t is the number of documents in the TREC set containing the term t and N refers to the size of the TREC set. Before computing TF-IDF scores (sum of TF-IDF values), we removed stop words but did not perform word stemming within MEDLINE documents. Moreover, in addition to text words, we also included MeSH® terms (obtained from PubMed's Automatic Term Mapping; see below for details) during the calculation of TF-IDF values. The final ranking of all of the retrieved documents depended upon their corresponding TF-IDF scores. A document with a higher TF-IDF score would be returned earlier in a list.

Rank by Sentence-level Co-occurrence

For each selected user query, we first obtained its translated form through PubMed's Automatic Term Mapping, which compares terms from the query with lists of terms comprising MeSH (including UMLS[27] mappings), journal titles, and author names. If a query term is untagged (without any search field tag) and maps to a MeSH term, the term will be searched as the MeSH term as well as in the Text Word field. For example, a search for *human parvovirus* in PubMed is automatically translated to:

> (("humans"[TIAB] NOT Medline[SB]) OR "humans"[MeSH Terms] OR human[Text Word]) AND ("parvovirus"[MeSH Terms] OR parvovirus[Text Word]) [a]

We ignored the search tag [TIAB] NOT Medline[SB] because it is designed to augment PubMed retrieval with additional non-MEDLINE citations that are out of scope of our analysis. For a query word (e.g., *parvovirus*) tagged with [MeSH Terms], we searched against all citations assigned with the MeSH term (i.e., parvovirus) as well as its more specific forms (e.g., H-1 parvovirus) in MeSH. For a query word (e.g., *human*) tagged with [Text Word], we performed an exact match against words in the abstract and title of a citation (i.e., human), as well as partial match to MeSH terms (e.g., Hepatitis, Viral, Human). Therefore, for each search word in a query, we were able to identify its occurrence(s) in Title, Abstract Sentence, and MeSH.

Next, we associated each retrieved document with one of the eight relevance levels defined in Relemed[18] (Table 2) during the course of identifying relationships of search words in multi-word user queries. Take the topic 229 for instance, we found eight documents (e.g., PMID 9192791) where the query words (i.e., *human* and *parvovirus*) occurred together in all three locations (Title, Same Abstract Sentence, and MeSH). Therefore, these eight documents were associated with the first relevance level. We processed all 3163 retrieved documents (see Table 2) for the 34 user queries in the same manner. The third column in Table 2 shows that co-occurrence existed in 67.7% of retrieved documents but only a small fraction of the retrieved documents were associated with the top relevance levels. For instance, over three quarters of the retrieved documents (77.6%) were associated with the last two relevance levels.

Finally, for each topic, the ranked list was assembled in the order of relevance levels. For those documents that were associated with the same relevance level, we ranked them in the reverse chronological order.

Evaluation Metrics

Many different measures for evaluating the performance of IR systems have been proposed[1], two of which are selected in this study: *mean average precision* and *mean rank precision*.

Mean Average Precision

The mean average precision is the mean value of the average precisions computed for all queries in our study. *Average precision* is the sum of the precision at each relevant document in the result set divided by the total number of relevant documents in the collection as shown below:

$$\text{Average precision} = \frac{\sum_{i=1}^{n}\left(precision(i) \cdot rel(i)\right)}{\text{number of relevant documents}} \quad (3)$$

where n is the number of returned documents; *precision(i)* is the precision at rank i; *rel(i)* is a binary function: at a given rank i, it equals 1 if the corresponding document is relevant, 0 otherwise. Average precision emphasizes returning more relevant documents earlier. Furthermore, to obtain a perfect

Table 3 ■ Results of Relevance Ranking on the 2007 TREC Topics by Different Strategies as Measured by Mean Average Precision (MAP), as well as Mean Precisions at Ranks 5, 10 and 20

Rank Strategy	MAP	Top 5	Top 10	Top 20
Reverse chronological order	0.126	0.385	0.411	0.413
TF-IDF weighting	0.182	0.538	0.525	0.503
Sentence-level co-occurrence	0.163	0.556	0.525	0.475
Perfect rankings	0.291	0.862	0.799	0.682

Under Perfect Rank, all relevant documents are ranked at the top. TF-IDF = term frequency–inverse document frequency; TREC = Text Retrieval Conference.

[a]Translations shown here were obtained in March 2008. Changes to PubMed after March 2008 may result in different translations.

Table 4 ■ Results of Relevance Ranking on the 2006 TREC Topics by Different Strategies as Measured by Mean Average Precision (MAP), as well as Mean Precisions at Ranks 5, 10 and 20

Rank Strategy	MAP	Top 5	Top 10	Top 20
Reverse chronological order	0.247	0.556	0.540	0.540
TF-IDF weighting	0.279	0.622	0.616	0.621
Sentence-level co-occurrence	0.270	0.575	0.573	0.583
Perfect rankings	0.405	0.908	0.830	0.761

Under Perfect Rank, all relevant documents are ranked at the top. TF-IDF = term frequency–inverse document frequency; TREC = Text Retrieval Conference.

average precision, all relevant documents need to be retrieved. Thus, this measure takes account of precision, relevance ranking, and overall recall.

Mean Rank Precision

The mean rank precision at rank *i* is the mean value of the rank precisions (shown as the value *precision(i)* in Formula 3) computed over all queries. In this study, we chose the cut-off ranks to be 5, 10, and 20 as most of the retrievals occur before 20 based on our own experience with PubMed users (data not shown here due to space constraints). Thus, the mean rank precision reveals more directly how a ranking strategy affects user retrieval effectiveness in practice.

Status Report

Experimental Results on 2007 Data

As can be seen from results of Mean Average Precision in the second column of Table 3, both relevance ranking strategies achieved better performance compared with the baseline performance. In addition, the comparison between the two relevance ranking strategies suggests that the ranking strategy dependent on the TF-IDF weighting is superior to its counterpart based on the sentence-level co-occurrence. The last row in Table 3 shows the best MAP one can possibly achieve given the current set of retrieved documents. Note that the best MAP is not a perfect 1 because not all relevant documents satisfy their respective queries.

Results of Mean Rank Precision in the last three columns of Table 3 suggest that the two relevance ranking strategies are equally successful considering the number of retrieved relevant documents at given ranks (5, 10, and 20). Furthermore, the difference between the two relevance ranking strategies and the baseline strategy is as small as a single relevant document in the top 5, 10, and 20 retrievals. Despite a small difference overall, retrieval efficiency can still be significantly different for individual topics. For instance, the rank precision increased substantially from 0.20 to 0.45 in the top 20 retrievals for Topic 235 when the baseline strategy was replaced by TF-IDF weighting. The best possible mean rank precisions are displayed in the last row of Table 3.

Table 5 ■ Mean Average Precisions are Compared with Regard to the Role of MeSH Terms in the TF-IDF Method in both TREC 2006 and 2007 Data Sets, Respectively

Data Set	Without MeSH Terms	With MeSH Terms
TREC 2006	0.279	0.279
TREC 2007	0.188	0.182

TF-IDF = term frequency–inverse document frequency; TREC = Text Retrieval Conference.

Table 6 ■ Mean Average Precisions are Compared with Regard to the Role of MeSH Terms in the Sentence-level Co-occurrence Method in Both TREC 2006 and 2007 Data Sets, Respectively

Data Set	Without MeSH Terms	With MeSH Terms
TREC 2006	0.251	0.270
TREC 2007	0.149	0.163

TREC = Text Retrieval Conference.

In order to compare results of different strategies statistically, we performed the bootstrap shift precision test at the 5% significant level.[28,29] We performed pair-wise comparisons for the results in Table 3. Results of these statistical tests[†] suggest that:

1. In terms of mean average precision, both relevance ranking strategies performed significantly better than the baseline strategy. Furthermore, there is a statistically significant difference between performance achieved by using TF-IDF weighting and sentence-level co-occurrence.
2. In terms of mean rank precision, both relevance ranking strategies performed significantly better than the baseline strategy but there is no statistically significant difference between performance achieved by the two relevance ranking strategies.

Experimental Results on 2006 Data

In addition to the experiments on the 36 topics in TREC 2007, we performed similar analyses on the 28 topics in TREC 2006 since the same text collection was used in both years. A total of 7 topics were removed before applying relevance ranking strategies: two were discarded because no relevant documents were found in the TREC set; the other five were excluded because our approach failed to generate word combinations to represent meaningful user queries (i.e., their F-measures are almost zero). For the remaining 21 topics, the mean and median query lengths are 2.6 and 2 words, respectively.

Results in Table 4 confirmed our previous observations of the retrieval effectiveness of relevance ranking strategies. Discrepancies between the results of 2006 and 2007 could be mostly attributed to the differences in topics and in the quality of relevance judgments.

Discussion

Role of MeSH Terms

A unique characteristic of PubMed searches (as opposed to Web searches in general) is the use of MeSH terms as we described earlier. Thus, in our experiments comparing the TF-IDF weighting and sentence-level co-occurrence ap-

[†]Detailed description of our statistical test is given as supplementary material, along with all of the pair-wise comparison results, publicly available at http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Relevance-ranking/supplementary.pdf.

Table 7 ■ Comparing Retrieval Effectiveness of Different Relevance Levels in the Sentence-level co-occurrence Strategy

Relevance level	Retrieved	Relevant	Precision
L1 T and A and M	18	12	0.67
L2 T and A	69	42	0.61
L3 T and M	16	5	0.31
L4 A and M	48	24	0.50
L5 T	57	32	0.56
L6 A	501	191	0.38
L7 M	1433	167	0.12
L8 T or A or M	1021	188	0.18
Total	3163	661	—

The second and third columns show the counts of retrieved and relevant documents, respectively. The computed precisions in the last column compare the retrieval effectiveness between different relevance levels. Same abbreviations used as in Table 2.

proach, MeSH terms play a role in both methods. We quantify their contribution in Tables 5 and 6, respectively. As can be seen, MeSH terms played a much more significant role in sentence-level co-occurrence than in TF-IDF.

Comparing Retrieval Effectiveness Between Different Relevance Levels

The strategy based on sentence-level co-occurrence defines eight relevance levels. As can be seen in Table 7, retrieved documents in relevance levels 1 to 6 are much more likely to be relevant than those in the last two levels. However, as we show earlier in Table 2, less than one quarter of the retrieved documents are associated with relevance levels 1 to 6. Thus, this limited the performance of the sentence-level co-occurrence method in our investigation.

Conclusions and Future Work

Based on the results of our large-scale analysis comprised of the 55 real biological questions and independently judged relevant documents, we conclude that TF-IDF weighting is the most effective strategy among the ones we examined here, and that the newly proposed sentence-level co-occurrence can deliver better performance than the baseline, but not as much improvement as TF-IDF. Therefore, we recommend TF-IDF weighting for relevance ranking in operational search engines like PubMed.

The comparison results are useful in suggesting changes in PubMed. However, as we pointed out earlier, this work makes two assumptions. One assumption is that the automatically generated queries represent real user queries. The other assumption is that the documents in the relevant set are the ground truth and there are no more relevant documents in the TREC set. Our future research goal is to address both issues by involving human experts in evaluations.

References ■

1. Hersh W. Information Retrieval: A Health and Biomedical Perspective. Spring-Verlag, 2nd edition, 2003.
2. Salton G. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
3. Salton G. Developments in automatic text retrieval. Science 1991;253(5023):974.
4. Salton G, Buckley C. Term weighting approaches in automatic text retrieval. Inf Process Manag. 1988;24:513–23.
5. Robertson SE, Walker S, Jones S, Hancock-Beaulieu M, Gatford M. Okapi at TREC-3. In Proceedings of the 3rd Text REtrieval Conference (TREC-3), 1994.
6. Zobel J, Moffat A. Exploring the similarity space. ACM SIGIR Forum 1998;32(1):18–34.
7. Keen EM. The use of term position devices in ranked output experiments. J Doc. 1991;47(1):1–22.
8. Keen EM. Some aspects of proximity searching in text retrieval systems. J Inf Sci. 1992;(18):89–98.
9. Clarke CLA, Gormack GV, Burkowski FJ. Shortest substring ranking. In Proceedings of the Fourth Text REtrieval Conference (TREC-4). 1995, pp. 294–304.
10. Hawking D, Thistlewaite P. Proximity operators—so near and yet so far. In Proceedings of the Fourth Text REtrieval Conference (TREC-4). 1995, pp. 131–43.
11. Tao T, Zhai CX. An exploration of proximity measures in information retrieval. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007; pp. 295–302.
12. Hatcher E, Gospodnetic O. Lucene in Action (In Action series). Manning Publications, 2004.
13. Cohen AM, Bhupatiraju RT, Hersh WR. Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. In Proceedings of the Thirteenth Text REtrieval Conference, 2004.
14. Carpenter B. Phrasal queries with LingPipe and Lucene: ad hoc genomics text retrieval. In Proceedings of the Thirteenth Text REtrieval Conference, 2004.
15. Caporaso GJ, Baumgartner WA, Cohen BK, Johnson HL, Paquette J, Hunter L. Concept recognition and the TREC genomics tasks. In Proceedings of the Fourteenth Text REtrieval Conference, 2005.
16. Hersh W, Bhupatiraju RT, Ross L, Johnson P, Cohen A, Kraemer D. TREC 2004 genomics track overview. In Proceedings of the Thirteenth Text REtrieval Conference, 2004.
17. Hersh WR, Cohen AM, Yang J, Bhupatiraju RT, Roberts P, Hearst M. TREC 2005 genomics track overview. In Proceedings of the Fourteenth Text REtrieval Conference, 2005.
18. Siadaty MS, Shu J, Knaus WA. Relemed: sentence-level search engine with relevance score for the MEDLINE database of biomedical articles. BMC Med Inform Decis Mak. 2007;7:1.
19. Hersh W, Cohen AM, Roberts P, Rekapalli HK. TREC 2007 genomics track overview. In Proceedings of the Sixteenth Text REtrieval Conference, 2007.
20. Hersh W, Cohen AM, Roberts P, Rekapalli HK. TREC 2006 genomics track overview. In Proceedings of the Fifteenth Text REtrieval Conference, 2006.
21. Wilbur WJ, Sirotkin K. The automatic identification of stop words. J Inf Sci. 1991;18(1992):45–55.
22. Wilbur WJ, Kim W, Xie N. Spelling correction in the PubMed search engine. Inf Retr Boston. Nov 2006;9(5):543–64.
23. Lu Z, Cohen KB, Hunter L. GeneRIF quality assurance as summary revision. Pac Symp Biocomput. 2007:269–80.
24. Kim W, Wilbur WJ. A strategy for assigning new concepts in the MEDLINE database, AMIA Annu Symp Proc. 2005:395–9.
25. Kim W, Aronson AR, Wilbur WJ. Automatic MeSH term assignment and quality assessment. AMIA Annu Symp Proc. 2001:319–23.
26. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinform 2007;8:423.
27. Lindberg D, Humphreys B, Mccray A. The unified medical language system. Methods Inf Med. 1993;32(4):281–91.
28. Noreen E. Computer Intensive Methods for Testing Hypotheses. John Wiley and Sons, 1989.
29. Wilbur W. Non-parametric significance tests of retrieval performance comparisons. J Inf Sci. 1994;20(4):270–84.