Published in final edited form as:
Int J Med Inform. 2011 June ; 80(6): 431–441. doi:10.1016/j.ijmedinf.2011.02.008.

MEDRank: using graph-based concept ranking to index biomedical texts

Jorge R. Herskovic, MD, PhD[1], **Trevor Cohen, MBChB, PhD**[1], **Devika Subramanian, PhD**[2], **M. Sriram Iyengar, PhD**[1,3], **Jack W. Smith, MD, PhD**[1], and **Elmer V. Bernstam, MD, MSE, MS**[1,4]

[1] School of Biomedical Informatics, The University of Texas Health Science Center at Houston

[2] Rice University Engineering School, Department of Computer Science

[3] NASA Johnson Space Center

[4] Department of Internal Medicine, Medical School, The University of Texas Health Science Center at Houston

Abstract

BACKGROUND—As the volume of biomedical text increases exponentially, automatic indexing becomes increasingly important. However, existing approaches do not distinguish central (or core) concepts from concepts that were mentioned in passing. We focus on the problem of indexing MEDLINE records, a process that is currently performed by highly-trained humans at the National Library of Medicine (NLM). NLM indexers are assisted by a system called the Medical Text Indexer (MTI) that suggests candidate indexing terms.

OBJECTIVE—To improve the ability of MTI to select the core terms in MEDLINE abstracts. These core concepts are deemed to be most important and are designated as “major headings” by MEDLINE indexers. We introduce and evaluate a graph-based indexing methodology called MEDRank that generates concept graphs from biomedical text and then ranks the concepts within these graphs to identify the most important ones.

METHODS—We insert a MEDRank step into the MTI and compare MTI’s output with and without MEDRank to the MEDLINE indexers’ selected terms for a sample of 11,803 PubMed Central articles. We also tested whether human raters prefer terms generated by the MEDLINE indexers, MTI without MEDRank, and MTI with MEDRank for a sample of 36 PubMed Central articles.

RESULTS—MEDRank improved recall of major headings designated by 30% over MTI without MEDRank (0.489 vs 0.376). Overall recall was only slightly (6.5%) higher (0.490 vs 0.460) as

Corresponding author: Elmer Bernstam, MD, Professor, School of Biomedical Informatics, University of Texas Health Science Center at Houston, 7000 Fannin, Suite 600, Houston, TX 77030, USA, Phone: 713 500 3901, Fax: 713 500 3929, elmer.v.bernstam@uth.tmc.edu.

Authors’ contributions: Drs. Bernstam and Herskovic participated in every phase of the work described in this manuscript. Dr. Cohen created the database of associations used by MEDRank. Drs. Subramanian, Iyengar and Smith participated in the formulation of the core ideas and drafting of the manuscript. Each co-author participated in data analysis and manuscript preparation. All coauthors have approved the final manuscript.

Statement on conflict of interest: None known.

was F_2 (3%, 0.408 vs 0.396). However, overall precision was 3.9% lower (0.268 vs 0.279). Human raters preferred terms generated by MTI with MEDRank over terms generated by MTI without MEDRank (by an average of 1.00 more term per article), and preferred terms generated by MTI with MEDRank and the MEDLINE indexers at the same rate.

CONCLUSIONS—The addition of MEDRank to MTI significantly improved the retrieval of core concepts in MEDLINE abstracts and more closely matched human expectations compared to MTI without MEDRank. In addition, MEDRank slightly improved overall recall and F_2.

Keywords (MeSH)

MEDLINE; PubMed; Digital Libraries; Abstracting and Indexing as Topic; Medical Informatics; Algorithms; Automatic Data Processing; Natural Language Processing

1. Introduction

Indexing is “the task of assigning to a document a *limited* number of terms denoting *concepts* that are *substantively* discussed in the document.” [1] The index terms “describe a document [and] serve as a synopsis of the subject matter discussed in the document.” [1] Index terms are useful for a variety of purposes including retrieving documents from a collection, identifying patients with particular diagnoses within a clinical data warehouse and summarizing documents. Although this paper is focused on indexing MEDLINE articles, we introduce a general approach that can, in principle, be used to index many types of documents.

MEDLINE is the premier collection of biomedical articles. It currently contains over 19,000,000 references from more than 5,000 biomedical journals, and grows continuously, at an ever-increasing rate. Finding relevant articles within MEDLINE requires good search tools and high-quality indexing that describes the content precisely. All MEDLINE entries corresponding to journal articles are indexed by hand using a purpose-built and continually maintained vocabulary called the Medical Subject Headings (MeSH). MeSH is a controlled, hierarchical vocabulary developed at the U.S. National Library of Medicine (NLM) and updated yearly [2]. MeSH terms are assigned to every article in MEDLINE by professional indexers. Thus, indexers must manually process approximately 670,000 articles per year [3] and assign MeSH subject headings (i.e., MeSH terms denoting concepts substantively discussed in the document) and major headings (i.e., MeSH terms reflecting the most important, central concepts discussed in the document) for each entry. Human indexing is expensive and inconsistent. Funk and Reid’s classic 1983 study of inter-indexer agreement showed that different indexers assigned the same MeSH terms to the same article between 33.8% (when comparing detailed concepts) and 74.7% (when comparing very general checktags) of the time [4]. Thus, automated indexing is an attractive alternative.

The NLM’s Medical Text Indexer (MTI) is a leading automated indexing effort in biomedicine. Unfortunately, MTI does not yet perform well enough to replace human indexers. Instead, it is used to suggest terms to human indexers [1, 5]. One reason for MTI’s inability to replace human indexers may be that MTI extracts concepts from article titles and abstracts rather than the full text [5]. As part of its processing MTI supplements the title and abstract with human-indexed concepts from similar articles [6, 7]. Current research on MTI focuses on subheading/heading attachment [1]. However, main heading selection is still critically important [8]. For a deeper look at MTI and other automated indexing initiatives see [8].

In this paper we present a novel way of indexing biomedical text called MEDRank that uses graph-based ranking algorithms. MEDRank operates on concepts extracted from text to

identify the most important terms and is thus complementary to indexing systems such as MTI.

2. Background

2.1 The structure of scientific writing

Frederick Suppe argued that scientific articles are rigidly structured [9]. Since journal space is a scarce resource, scientific articles must use available space optimally to advance their claims. An intuitive consequence of this theory is that scientific papers build a network of interrelated concepts. Authors advance their claims by stating facts about those claims, about related concepts or the relationships between concepts. These concepts and their inter-relationships form a network that reflects the concepts in the original text. Thus, the most "important" or "central" concepts in the network will be the most important concepts in the text.

We leverage the structure of scientific writing by creating graphs that represent the concepts in biomedical articles. We then apply graph-based ranking algorithms to identify the "most important" nodes, i.e. the most important nodes as given by their relationships to other nodes. We hypothesize that these "most important nodes" correspond to the most important concepts in the articles, and are therefore good indexing terms. In particular, the highest-ranked concepts should correspond to the major headings chosen by human indexers.

The approach we propose is also consistent with Kintsch's widely used construction-integration model of text comprehension [10], in which concepts occurring in a passage of text, and related concepts from the memory of the reader, form an associative network in the mind. Spreading activation across this network causes concepts that are most highly connected to dominate the cognitive representation of the passage.

2.2 Graph-based ranking algorithms

A graph is "a diagram consisting of a set of points together with lines joining certain pairs of these points" [11]. The points and lines in graph theory are commonly called "nodes" and "edges" respectively. In this paper, we use Semantic Abstraction Graphs (SAGs) which represent concepts from a piece of text as nodes and relationships between these concepts as edges in the graph [12].

Graph theory and graph analysis are useful in dealing with many kinds of human-created networks. Perhaps the best-known example of graph analysis is Google (http://www.google.com). Google models the Web as a graph. Web pages are represented as nodes, and the hyperlinks are the edges. The graph is analyzed using an algorithm called PageRank. PageRank has been applied to multiple networks, including biomedical literature citation networks [13], social networks [14], and text to achieve summarization by selecting important sentences [15]. PageRank is thus a general algorithm that will rank nodes in a graph based on their relative importance as established by the set of edges. TextRank is a variant of PageRank that was created specifically to work on undirected graphs with weighted links and performs well when choosing keywords out of text [15]. MEDRank uses PageRank (for directed graphs) as described in [16] or TextRank (for undirected graphs) as described in [15] modified to operate on SAGs derived from text documents as opposed to web pages and text words respectively.

2.3 MEDRank overview

The basic architecture of MEDRank (available as Open Source code from http://github.com/drh-uth/MEDRank) is shown in Figure 1. First, we split documents into

individual sentences. Then, we feed each sentence separately to a concept extraction stage that returns an ordered list of concepts for each sentence. MEDRank can use the list of concepts to infer relationships between them, or it can accept a list of relationships between concepts. MEDRank uses the concepts and relationships to generate SAGs. It then ranks the concepts in the SAG with a graph-based ranking algorithm. Finally, MEDRank translates the ranked list of concepts into the destination indexing vocabulary; in this case, MeSH.

3. Methods

3.1 Rationale

To verify that MEDRank can help MTI discover the most important indexing terms in a biomedical article, we compared MTI with and without MEDRank. MTI processing is complex [17–20] and contains several steps that are difficult to replicate exactly, including the Restrict to MeSH algorithm [19], and a clustering/ranking step [20]. To evaluate the contribution of MEDRank, we used a modified version of MTI. NLM staff prepared this modified version of MTI at our request. It separates MetaMap from the later stages MTI processing, allowing us to inject MEDRank's processing into MTI's pipeline (Figure 2). In summary, we processed articles using MetaMap and MTI but added a stage where the extracted UMLS concepts were ranked with MEDRank. We call this MTI+MEDRank.

Additionally, we know that indexers review the abstract and title of an article carefully and skim the text of the paper [21]. Therefore, we compared MTI to MTI+MEDRank using only the abstract and title.

In summary, MTI+MEDRank differs from MTI in two important ways. First, MTI +MEDRank uses graph-based ranking, while MTI does not. Second, MTI's normal workflow uses MetaMap to extract concepts from the abstract and title of articles as a whole, while MTI+MEDRank processes each sentence separately. This change may affect MTI's performance. Therefore we evaluated it separately.

We recognize that there is no single definitive strategy for evaluating indexing systems. We chose to use the terms chosen by NLM (human) indexers as the gold standard. As noted above human indexers are not always consistent with an agreement ranging between 74.7% for checktags to 33.8% for heading/subheading combinations [4]. Checktags, such as HUMAN, ANIMAL, MALE and FEMALE, are "large-volume descriptors routinely checked for in every indexed article."[22] Since they are very general, checktags are not good indicators of what the article is really about. In contrast, major headings "reflect the central concepts of an article." [4] Further, inter-indexer agreement was relatively good for major headings (61.1%). While evaluation of overall recall and precision provides a measure of the match between system and indexer preference, it does not address the relative importance of selected indexing terms. Consequently, while we have reported overall statistics, the emphasis of our evaluation is on the recall of major headings. Finally, we expect the system to have a measurable impact in the real world. We therefore determined whether blinded human readers preferred terms produced by MEDLINE indexers, MTI, or MTI+MEDRank (see below for details).

3.2 MetaMap

We used the NLM's MetaMap program to extract concepts from the article's abstract and title. We ran MetaMap directly on the NLM's Semantic Knowledge Representation (SKR) server, available at http://skr.nlm.nih.gov, using the Batch Generic with Validation facility and the command "metamap0809 –iDNE" as suggested by NLM staff to emulate the first stage of MTI processing. MetaMap produces a list of UMLS concepts and a confidence score between 0 and 1000 for each concept.

3.3 MEDRank

The pseudocode for MEDRank is given in Figure 3, below.

We performed two experiments with different SAGs. We first created SAGs by inferring relationships between adjacent concepts where the distance between concepts determined the strength of the relationship. We set the initial score of the nodes in our SAGs to the normalized MetaMap confidence score. We set the weight of the relationship between two concepts to $\frac{1}{e^d}$ where d was the number of concepts between them. In other words, we considered concepts to be closely related if they appeared close to each other in the text; the strength of the relationship decayed exponentially with distance. To speed processing and prune weak links we arbitrarily rounded the weight of any relationship with a d of 5 or more to 0. We also created SAGs by using a database of known relationships between UMLS concepts produced by Reflective Random Indexing (described below). As all SAGs in these experiments are undirected, we used our modified TextRank to rank the nodes.

3.4 Random Reflective Indexing

Distributional models of semantic relatedness correlate well with human estimates in general [23], as well as in the biomedical domain [24]. Consequently, we chose to utilize a variant of the Random Indexing [25, 26] method to measure the semantic relatedness between UMLS concepts in order to weight the links of our SAGs. We derive these measurements from a corpus of all abstracts and titles that have been added to MEDLINE over the past decade.

We generated semantic vectors [27] for each UMLS concept in the data set. These semantic vectors are derived from the distribution of terms that occur in documents with each concept. We then measured associations between UMLS concepts using the normalized dot product between vectors (i.e. a vector cosine comparison). The advantage of using a semantic vector representation is that its reduced dimensionality makes it small enough to retain in RAM, and therefore it is possible to rapidly calculate associations between concepts that occur in each SAG (please see [8, 23, 25, 26, 28] for more information on these techniques).

3.5 MTI

NLM staff provided a modified MTI that accepts a custom delimited input format. This “custom” version of MTI allowed us to split the normal MTI processing of articles [29, 30] into a concept extraction phase (performed by running MetaMap as described above) followed by ranking and filtering. In our experiments, MTI performed the clustering, converting to MeSH, and ranking steps that are part of the regular MTI workflow. We left the output length at the MTI default of 25 terms. Therefore, running this modified MTI on the output of the MetaMap phase described above produced the same results as running regular MTI on the input text.

3.6 Sample

We used a custom Python script to retrieve a list of all articles in PubMed Central (ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/file_list.txt), and repeated the following procedure until 14,000 distinct, randomly-selected articles were downloaded. The script downloaded random articles and verified that each article was associated with a PubMed record containing MeSH indexing terms. To facilitate future indexing experiments on full text, we chose articles that were stored in PubMed Central. We used the 2008 edition of MeSH to perform our experiments. Therefore, we excluded articles without MeSH 2008 mappings from the sample (see “Results” below).

3.7 Indexing Experiments

We first indexed the articles' title and abstract using the MTI processing pipeline. In other words, we used MetaMap configured as described above, on the entire abstract and title at once as is usual for MTI processing, plus our custom MTI.

We then indexed the articles' titles and abstracts by extracting the concepts sentence by sentence using MetaMap separately on each sentence. We processed this list of concepts with MTI to determine whether the different extraction process affects performance. We found that MetaMap identifies more concepts when extracting sentence by sentence than when extracting from the entire abstract and title simultaneously. Thus, we expected MTI's recall to be higher and its' precision to be lower when using sentence-by-sentence extraction. We also used these concepts extracted sentence by sentence to build the SAG for each article, which we then ranked using MEDRank. We performed the experiment configuring MEDRank to discard all concepts with a ranking score of 0.50 or less, a threshold we determined empirically during development. We rescaled MEDRank scores, normally between 0 and 1, to a scale between 0 and 1000 to imitate MetaMap output. The output from MEDRank became the input to the rest of MTI processing (see Figure 2). Consequently, in addition to MetaMap's confidence score, the modified MTI+MEDRank pipeline takes into account the extent to which each extracted concept relates to other concepts that occur in the citation concerned, providing a measure of sensitivity to context.

The goal of our experiment was to determine whether MTI+MEDRank could identify the most important concepts in biomedical articles. Major headings are thought to represent the most important concepts in an article. Thus, our main outcome measure was the micro-averaged recall (the average across the entire sample of the recall on each document) of major headings from the MEDLINE record. We defined recall of major headings as the number of MeSH terms marked by an indexer as a major heading present in MTI's output, since MTI does not mark terms as major headings. We also report micro-averaged recall, precision, and F_2 (the harmonic mean of precision, recall, and recall) for all MeSH headings. We also report major heading precision at ranks 5, 10, 15, 20, and 25. As the major headings are embedded in the overall retrieval process we also report the maximum achievable precision at a certain rank, computed by assuming that all major headings are ranked before any other terms.

We analyzed results using R 2.11.1 (http://r-project.org) running on Mac OS X 10.6.4 (Apple Computer, Inc., Cupertino, CA). Since major heading recall was not normally distributed, we used a paired non-parametric Wilcoxon rank-sum test to compare results. We also estimated the mean and 95% confidence interval using a basic non-parametric bootstrap implemented by the R Hmisc library function smean.cl.boot.

3.8 Reader preference experiment

We hypothesized that identifying the "most important" concepts in an article would be a better match to searcher expectations than MTI. In other words, we asked the question: "Do indexing terms selected by MEDRank match searcher expectations better than MTI?" We therefore determined whether MEDLINE users prefer terms generated MTI+MEDRank or by MTI alone. We also compared the results to terms chosen by MEDLINE indexers.

Three authors with medical training (EVB, TC, and JRH) voted for five potential indexing terms for a random sample of articles. We were blinded to the source of the term (i.e., whether the term was selected by MTI, MTI+MEDRank or by MEDLINE indexers). In an informal experiment on three articles (15 terms chosen), we tried to identify the source of the term (i.e., which system generated the term) and there was no apparent correlation between our "guesses" and the actual source system. Thus, it is not likely that our results

were biased by knowledge of the study or the systems being evaluated. We simply could not tell which system generated the terms for a previously-unseen random article. We selected terms that best captured the content of the article. Each rater received a set of pages with an article's title and abstract, and a randomly ordered list of terms selected at random from the output of MTI, MTI+MEDRank, and the MEDLINE record for the article. Since a given term could have more than one source (for example, [myocardial infarction] could be "selected" by MTI, MTI+MEDRank and MEDLINE indexers) we sampled from the lists with replacement and randomly removed terms from the output until each source had the same number of potential votes. In other words, the prior probability of a vote going to a term from each source was identical. The authors did not communicate with each other while performing this experiment, and were blinded to the source of the terms (Figure) (i.e., the raters did not know whether the term was selected by MTI, MTI+MEDRank or by MEDLINE indexers).

Each rater received fifteen articles for the reader preference study. The first three of these randomly chosen articles were identical, and were used only to compare inter-rater consistency. We computed the average of three two-way Hooper's consistency measures for each article [4]. The other 12 articles for each author were different, and were used to compute reader preference. We asked each rater to "vote" for exactly five terms in each article. We then counted the votes as points for each source "system" that "selected" the term. Terms that were present in the output of two systems gave a point to each. We compared the number of votes the different systems obtained using the normal approximations of the paired Wilcoxon rank sum tests, which gave us an estimated median of the difference between groups and a p value. As we performed three comparisons (MEDLINE vs. MTI, MEDLINE vs. MTI+MEDRank, MTI+MEDRank vs. MTI) we used a Bonferroni correction to establish a desired alpha of 0.016667 (0.05 divided by 3). We chose the sample size of 36 articles to detect at least a 0.5 term/article preference difference with a statistical power of 0.8 and an alpha of 0.016667, using non-parametric statistics (we computed the power required for a t-test and overpowered the study by 15% [31]). In other words, we were able to detect a statistically significant difference in preferences of at least one term every two articles. We considered a difference of less than one term every two articles to be practically insignificant.

4. Results

Our final sample contained 11,803 articles after excluding articles without 2008 MeSH headings (which includes all articles published after MeSH 2009 was made available), and articles where MetaMap or MTI malfunctioned during processing. A total of 339 articles (2.5%) were excluded due to MetaMap or MTI malfunction. Notably, these 339 articles were excluded from both MTI and MTI+MEDRank groups.

4.1 MTI

On these articles the default MTI processing pipeline was able to retrieve major headings with a micro-averaged recall of 0.376 (95% CI: 0.370–0.381). Switching the concept extraction step to a sentence-by-sentence extraction improved major heading recall using the default MTI workflow to 0.405 (95% CI: 0.399–0.410) at the cost of decreased precision (Table 1).

4.2 MEDRank indexing experiments

When we added MEDRank to MTI (MEDRank+MTI), major heading recall improved by 26% to 0.475 (95% CI: 0.469–0.481) (Table 1). Therefore, MEDRank was able to improve MTI's recall of major headings significantly when working with just the titles and abstracts

(Wilcoxon paired sum-rank test, $p<0.001$). Recall of major headings improved by 30%% over MTI to 0.489 (95% CI: 0.484–0.495) when using Reflective Random Indexing (RRI) [26] instead of sentence-level co-occurrence to weight the edges of the SAGs. Overall F_2 was also modestly (3%) better when applying MEDRank+MTI using RRI to weight edges than when using MTI alone (0.408, 95% CI: 0.406–0.410 versus 0.396, 95% CI: 0.395–0.399). MEDRank also significantly improved MTI's precision of major heading retrieval at all ranks (Figure 5).

4.3 Reader preference study

In our preference study, we agreed on indexing terms on three articles reviewed by all three authors within the ranges reported by Funk and Reid [4]. The Hooper's indexing consistency values were 37%, 51%, and 37%. Raters significantly preferred terms generated by MEDLINE indexers over MTI (estimated median difference=1.00 votes/article, $p < 0.010$) and MEDRank+MTI over MTI (estimated median difference=1.00 votes/article, $p= 0.015$). There was no significant difference in reader preference between MEDRank+MTI and MEDLINE (estimated median difference=0.05 votes/article in favor of MEDLINE, $p=0.36$).

5. Discussion

We presented MEDRank, an application of graph-based ranking. We found that MEDRank can be "inserted" into the MTI pipeline and can improve recall of major headings by 30%, and the precision @ rank 5 of major headings by 47%. Further, MEDRank+MTI performed comparably to human MEDLINE indexers and significantly better than MTI alone when compared against human expectations of indexing terms associated with a particular article.

The addition of MEDRank, our graph-based algorithm designed to identify the most important concepts in a document, to MTI improves indexing performance. The use of sentence-by-sentence concept extraction also improved the recall of major headings. The combination of sentence-by-sentence concept extraction and MEDRank increased major heading recall from MTI's 0.376 to 0.475; a 26% relative increase. When we derived edge weights from a database of known co-occurrence between concepts drawn from a larger sample of the biomedical literature, recall of major headings increased again to 0.489, a 30% increase relative to MTI's baseline. It was previously suggested that most major gains in MTI performance had already been realized [32]. Our experience with MEDRank contradicts this and shows that new approaches based on novel theories can improve indexing performance. Further, our reader preference evaluation supports the idea that the terms MEDRank+MTI identifies are qualitatively 'better' (i.e., are a closer match to searcher expectations) than the terms MTI identifies, and may even be similar to human MEDLINE indexer performance.

MEDRank+MTI are also able to surpass MTI's general performance measured by F_2, albeit modestly, which was surprising. In contrast to MEDRank, MTI is actively optimized for F_2 (see [32]). F_2 is an appropriate measure for MTI's main goal, suggesting terms to the NLM indexers, but it is not ideal for measuring whether an algorithm identified the most important concepts in a biomedical article. Despite this limitation the addition of MEDRank to the MTI process improved MTI's performance (Table 1). As MEDRank can work within the existing NLM infrastructure, it could be adopted as-is and added to the indexing process.

Selecting appropriate descriptors, i.e., indexing terms for a document is a well-known problem in information retrieval [5, 8], and is of particular interest to biomedical information retrieval. Recent research in biomedical indexing focuses on improving the quality and quantity of subheadings and main heading/subheading pairing [1] and on improving main heading selection [8].

Similar graph-based algorithms have used to select keywords from free text [15]. Other graph-based algorithms have been used to cluster terms into semantically similar clusters [33]. To our knowledge, this is the first application of a graph-based ranking algorithm to a concept graph and the first implementation of a SAG for indexing purposes.

It is difficult to compare our quantitative results directly to previous work. Many previous studies of automated or semi-automated indexing focused on a particular domain (e.g., genetics) [34], did not focus on MEDLINE (e.g., dermatology atlas) [35] and/or used a very different gold standard (e.g., a 200-article test set [8]. Further, MTI changes over time and thus affects both MTI results and, to some extent, the terms chosen by human MEDLINE indexers. Thus, we have chosen to perform a relative comparison (MTI vs. MTI+MEDRank vs. human MEDLINE indexing) using a large, randomly selected subset of MEDLINE. Further, we also compared systems with respect to their match to searcher expectations.

Our study had several limitations. Perhaps most importantly, we cannot conclude that the addition of MEDRank to MTI indexing will improve the practical experience of MEDLINE users. To draw this conclusion, we would need a user study with representative MEDLINE users. However, we found that blinded raters preferred terms selected by MTI+MEDRank compared to MTI alone. Further, the difference between MTI+MEDRank and human MEDLINE indexers was not statistically significant. Thus, there is reason to believe that if articles were indexed using MTI+MEDRank, user experience may be comparable to the status quo that relies on human indexers.

MTI+MEDRank also showed a modest but significant decrease in precision when compared to MTI. It could thus be argued that all we do is trade off some precision for recall. Our task, however, was to establish whether MEDRank can identify the most important concepts in the article, which we operationalized as the starred major headings. We traded off a small 4% relative loss in precision (between baseline MTI and MTI+MEDRank using the relationship database) for a 30% relative increase in major indexing recall. The increase in general recall was large enough to offset the precision losses and obtain a net increase in F_2, the main MTI outcome measure. Finally, in our reader preference experiment, we show that readers select terms from MTI+MEDRank as much as from MEDLINE, suggesting a qualitative increase in term quality. Any evaluation of MTI using MEDLINE as the gold standard is inherently biased towards MTI. Since some MEDLINE indexers use MTI during their work it is likely that, if there are two equally appropriate terms, the indexer will choose the term suggested by MTI. Unfortunately, we are not aware of any objective way to measure this effect [36].

5.1. Future work

The work we present raises issues that we will explore in the future. We show that MEDRank matches end-users' expectations better than MTI. It is possible that MEDRank will also match indexers' expectations better than MTI. We will attempt to evaluate this hypothesis. We will also study the effect of MetaMap's newly-added support for negations [37] on MEDRank's performance, and continue working to improve MEDRank's quality.

6. Conclusion

MEDRank is a novel graph-based algorithm that can improve the performance of existing concept extraction systems. We found that adding MEDRank to MTI, the current state of the art in MEDLINE indexing, improved recall of important concepts by 30% at the cost of a slight (4%) decrease in overall precision. Further, MEDRank helped MTI select terms that are likely to match searcher expectations. Since MEDRank does not incorporate expert knowledge regarding a particular indexing tasks or specific vocabularies, we are applying

MEDRank to other indexing tasks including identifying important concepts within clinical text.

Acknowledgments

This work was partially supported by: NCRR Grant 3UL1RR024148, NLM Grant 5K22LM8306, and NCRR Grant 1RC1RR028254. The authors are also grateful for help and technical support from the Semantic Knowledge Representation and MetaMap teams at the National Library of Medicine, especially Mr. Jim Mork and Dr. Olivier Bodenreider.

Bibliography

1. Névéol A, Shooshan SE, Humphrey SM, Mork JG, Aronson AR. A recent advance in the automatic indexing of the biomedical literature. J Biomed Inform. 2008 Dec 30; 42(5):814–23. [PubMed: 19166973]
2. U.S. National Library of Medicine. MeSH history. [Web page]. Bethesda, MD: National Library of Medicine; 2006. [updated November 27; cited 2007 March 30]; Available from: http://www.nlm.nih.gov/mesh/intro_preface2007.html#pref_hist
3. U.S. National Library of Medicine. Medical Subject Heading (MESH) Fact Sheet. [Web page]. Bethesda, MD: U.S. National Library of Medicine; 1999. [updated October 30, 2007; cited 2008 September 25]; Available from: http://www.nlm.nih.gov/pubs/factsheets/mesh.html
4. Funk ME, Reid CA, McGoogan LS. Indexing Consistency in MEDLINE. Bull Med Libr Assoc. 1983; 71(2):176–83. [PubMed: 6344946]
5. Gay CW, Kayaalp M, Aronson AR. Semi-automatic indexing of full text biomedical articles. AMIA Annu Symp Proc. 2005:271–5. [PubMed: 16779044]
6. Kim W, Aronson AR, Wilbur WJ. Automatic MeSH term assignment and quality assessment. Proc AMIA Symp. 2001:319–23. [PubMed: 11825203]
7. U.S. National Library of Medicine. PubMed Related Citations algorithm. 2004. [updated March 16; cited 2009 December 29]; Available from: http://ii.nlm.nih.gov/MTI/related.shtml
8. Vasuki V, Cohen T. Reflective random indexing for semi-automatic indexing of the biomedical literature. J Biomed Inform. 2010 Apr 9; 43(2):240–56. [PubMed: 19761870]
9. Suppe F. The Structure of a Scientific Paper. Philos Sci. 1998; 65(3):381–405.
10. Kintsch, W. Comprehension: A Paradigm for Cognition. Cambridge, UK: Cambridge University Press; 1998.
11. Bondy, JA.; Murty, USR. Graph Theory with Applications. New York, NY: Elsevier Science Publishing Co; 1976.
12. Fiszman, M.; Rindflesch, TC.; Kilicoglu, H. Abstraction Summarization for Managing the Biomedical Research Literature. Proc HLT NAACL Workshop on Computational Lexical Semantics; 2004. p. 76-83.
13. Bernstam EV, Herskovic JR, Aphinyanaphongs Y, Aliferis CF, Sriram MG, Hersh WR. Using Citation Data to Improve Retrieval from MEDLINE. J Am Med Inform Assoc. 2006 Jan–Feb; 13(1):96–105. [PubMed: 16221938]
14. Pujol, JM.; Sanguesa, R.; Delgado, J. Extracting reputation in multi agent systems by means of social network topology. Proceedings of the 1st International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS); 2002. p. 467-74.
15. Mihalcea, R.; Tarau, P., editors. TextRank: Bringing Order into Texts; Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004); Barcelona, Spain: 2004 July.
16. Page, L.; Brin, S.; Motwani, R.; Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. Stanford Publications; 1998. [03/15/2005]; Available from: http://dbpubs.stanford.edu:8090/pub/1999-66
17. U.S. National Library of Medicine. MetaMap Indexing Algorithm. 2004. [updated March 16; cited 2008 November 6]; Available from: http://ii.nlm.nih.gov/MTI/mmi.shtml

18. U.S. National Library of Medicine. Trigram Algorithm. 2004. [updated March 16; cited 2008 November 6]; Available from: http://ii.nlm.nih.gov/MTI/trigram.shtml
19. U.S. National Library of Medicine. Restrict to MeSH algorithm. 2004 March 16. 2008(November 6)
20. U.S. National Library of Medicine. Clustering and Ranking process. 2004. [updated March 16; cited 2008 November 6]; Available from: http://ii.nlm.nih.gov/MTI/cluster.shtml
21. U.S. National Library of Medicine. Principles of MEDLINE Subject Indexing. [Web page]. Bethesda, MD: National Library of Medicine; 2007. [updated February 23; cited 2007 April 18]; Available from: http://www.nlm.nih.gov/bsd/disted/mesh/indexprinc.html
22. U.S. National Library of Medicine. Online services reference manual. Bethesda, MD: 1982.
23. Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. J Biomed Inform. 2009; 42(2):390–405. [PubMed: 19232399]
24. Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform. 2007; 40:288–99. [PubMed: 16875881]
25. Kanerva, P.; Kristofersson, J.; Holst, A., editors. Proceedings of the 22nd Annual Conference of the Cognitive Science Society. University of Pennsylvania; 2000. Random indexing of text samples for latent semantic analysis.
26. Cohen T, Schvaneveldt R, Widdows D. Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections. J Biomed Inform. 2010; 43(2):240–56. [PubMed: 19761870]
27. Hecht-Nielsen, R. Context vectors; general purpose approximate meaning representations self-organized from raw data. In: Zurada, JM.; RJM; Robinson, CJ., editors. Computational intelligence: imitating life. IEEE Press; 1994.
28. Sahlgren, M., editor. Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering. TKE; Citeseer: 2005. An introduction to random indexing.
29. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Studies in health technology and informatics. 2004; 107(Pt 1):268–72. [PubMed: 15360816]
30. U.S. National Library of Medicine. Medical Text Indexer (MTI). Bethesda, MD: 2009. [cited 2009 December 21]; Available from: http://ii.nlm.nih.gov/mti.shtml
31. Lehmann, Lehmann EL.; Nonparametrics, Erich L. With the special assistance of H J M D'Abrera. New York: Springer; 2006. Statistical methods based on ranks. Revised first edition
32. Gay, CW. Summary of Threshold Studies. U.S. National Library of Medicine, Lister Hill National Center for Biomedical Communications; 2006.
33. Ohsawa, Y.; Benson, N.; Yachida, M., editors. Proceedings of the Advanced Digital Library Conference. IEEE Computer Society; 1998. KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor.
34. Névéol, A.; Shooshan, S.; Humphrey, S.; Rindflesch, T.; Aronson, A., editors. Pac Symp Biocomput. 2007. Multiple approaches to fine-grained indexing of the biomedical literature.
35. Kim G, Aronson A, Mork J, Cohen B, Lehmann C. Application of a Medical Text Indexer to an online dermatology atlas. Stud Health Technol Inform. 2004; 107(Pt 1):287. [PubMed: 15360820]
36. Ruiz, ME.; Aronson, AR. User-centered Evaluation of the Medical Text Indexing (MTI) system. Vol. 2007. U.S. National Library of Medicine; 2007.
37. Aronson, AR.; Lang, F., editors. AMIA 2009. San Francisco, CA: AMIA; 2009 November 16. The Evolution of MetaMap, a Concept Search Program for Biomedical Text.

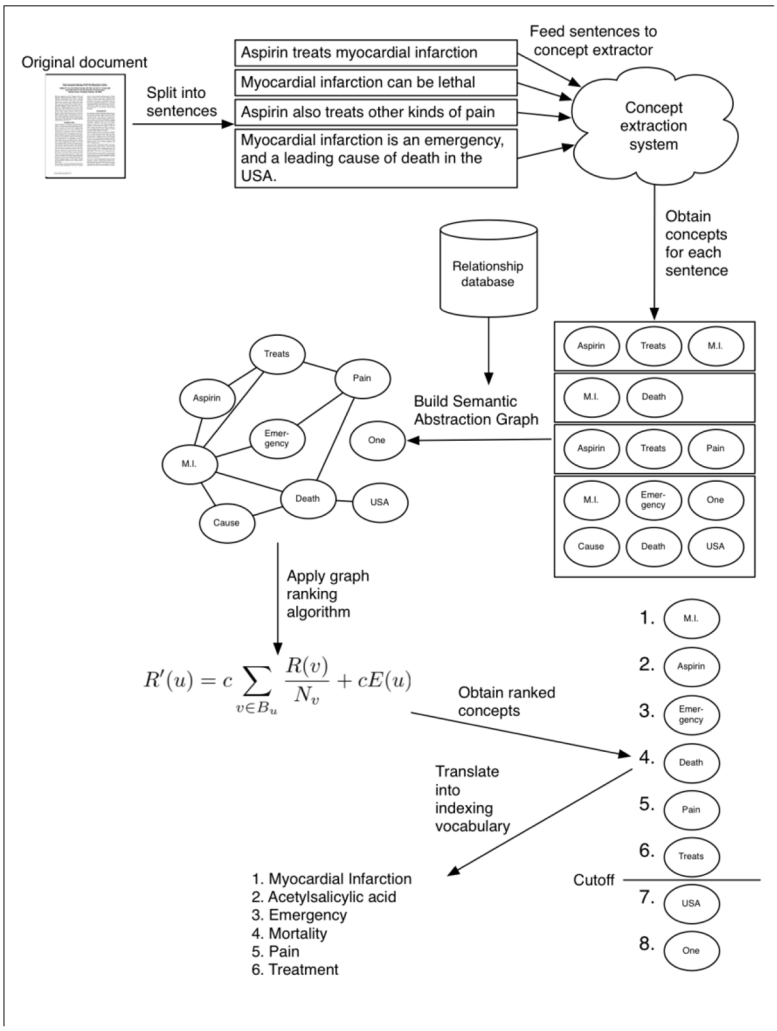


Figure 1.
MEDRank Standalone Workflow

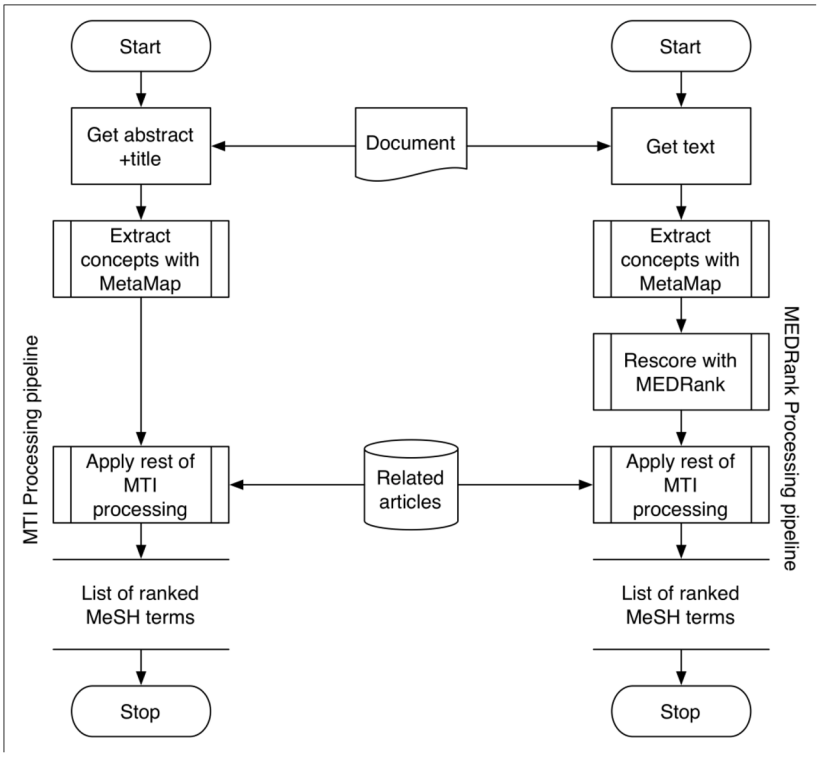


Figure 2.
MTI and MEDRank processing pipelines

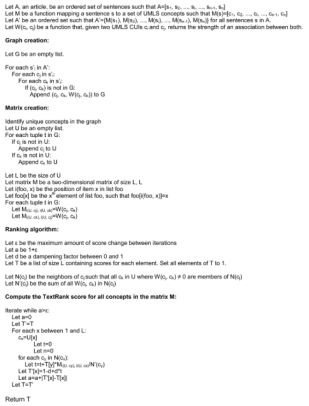
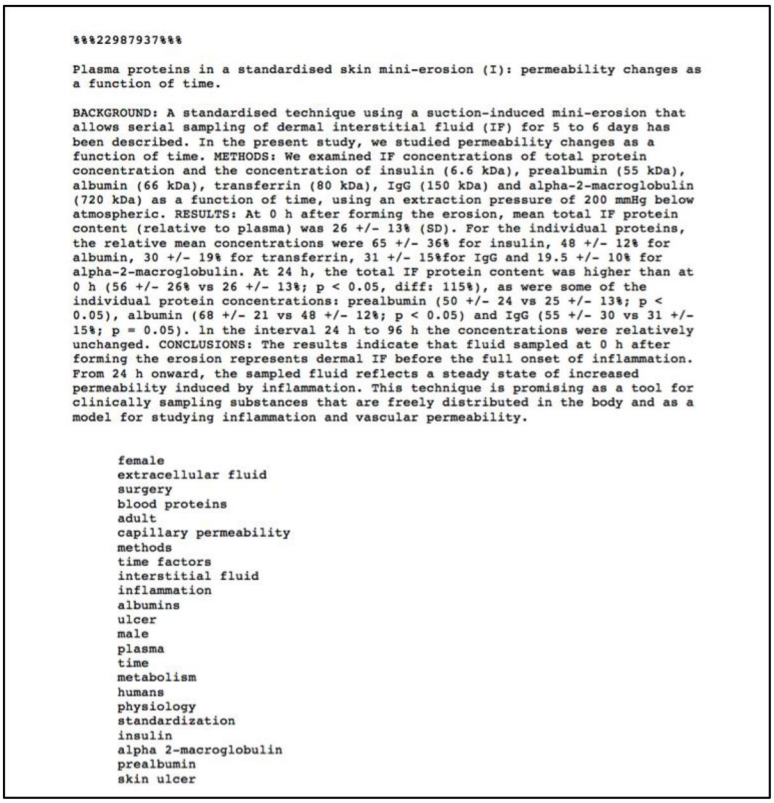


Figure 3.



%%%22987937%%%

Plasma proteins in a standardised skin mini-erosion (I): permeability changes as a function of time.

BACKGROUND: A standardised technique using a suction-induced mini-erosion that allows serial sampling of dermal interstitial fluid (IF) for 5 to 6 days has been described. In the present study, we studied permeability changes as a function of time. METHODS: We examined IF concentrations of total protein concentration and the concentration of insulin (6.6 kDa), prealbumin (55 kDa), albumin (66 kDa), transferrin (80 kDa), IgG (150 kDa) and alpha-2-macroglobulin (720 kDa) as a function of time, using an extraction pressure of 200 mmHg below atmospheric. RESULTS: At 0 h after forming the erosion, mean total IF protein content (relative to plasma) was 26 +/- 13% (SD). For the individual proteins, the relative mean concentrations were 65 +/- 36% for insulin, 48 +/- 12% for albumin, 30 +/- 19% for transferrin, 31 +/- 15%for IgG and 19.5 +/- 10% for alpha-2-macroglobulin. At 24 h, the total IF protein content was higher than at 0 h (56 +/- 26% vs 26 +/- 13%; p < 0.05, diff: 115%), as were some of the individual protein concentrations: prealbumin (50 +/- 24 vs 25 +/- 13%; p < 0.05), albumin (68 +/- 21 vs 48 +/- 12%; p < 0.05) and IgG (55 +/- 30 vs 31 +/- 15%; p = 0.05). In the interval 24 h to 96 h the concentrations were relatively unchanged. CONCLUSIONS: The results indicate that fluid sampled at 0 h after forming the erosion represents dermal IF before the full onset of inflammation. From 24 h onward, the sampled fluid reflects a steady state of increased permeability induced by inflammation. This technique is promising as a tool for clinically sampling substances that are freely distributed in the body and as a model for studying inflammation and vascular permeability.

female
extracellular fluid
surgery
blood proteins
adult
capillary permeability
methods
time factors
interstitial fluid
inflammation
albumins
ulcer
male
plasma
time
metabolism
humans
physiology
standardization
insulin
alpha 2-macroglobulin
prealbumin
skin ulcer

Figure 4.
Example rating page for readers. The first line contains a unique article ID.

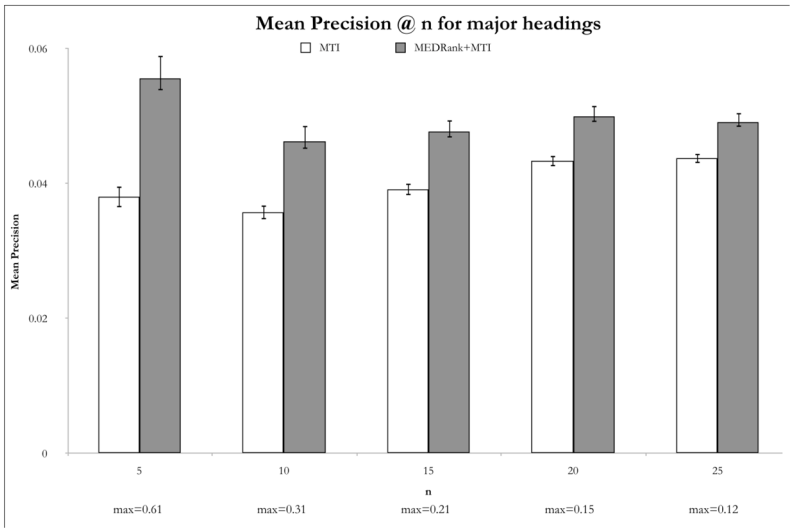


Figure 5.
Column graph showing bootstrapped micro-averaged mean precision at rank n and 95% confidence intervals for major headings for MTI and MEDRank+MTI using distributional relationships. We show the maximum achievable precision at each rank below the x axis.

Table 1

Performance measures: bootstrapped micro-averaged mean and 95% confidence interval by indexing technique (n=11,803) (best performance in bold)

	Recall (major headings)	Recall (all headings)	Precision (all headings)	F_2 (all headings)
Traditional MTI workflow (whole abstract and title)	0.376 [0.370–0.381]	0.460 [0.458–0.463]	**0.279 [0.278–0.281]**	0.396 [0.395–0.399]
MTI workflow on the abstract and title, extracted sentence by sentence	0.405 [0.399–0.410]	0.472 [0.469–0.474]	0.277 [0.276–0.279]	0.402 [0.400–0.404]
MEDRank + MTI on abstract and title (co-occurrence relationships)	0.475 [0.469–0.481]	0.461 [0.458–0.464]	0.248 [0.246–0.249]	0.381 [0.379–0.383]
MEDRank +MTI on abstract and title (distributional relationships derived using RRI)	**0.489 [0.484–0.495]**	**0.490 [0.487–0.492]**	0.268 [0.266–0.270]	**0.408 [0.406–0.410]**