Published in final edited form as:
Proc 1st Workshop Sch Web Min (2017). 2017 February ; 2017: 49–56. doi:10.1145/3057148.3057155.

Predicting MeSH Beyond MEDLINE

Adam K. Kehoe,
School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL

Vetle I. Torvik[*],
School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL

Matthew B. Ross, and
Department of Economics, Ohio State University, Columbus, OH

Neil R. Smalheiser
Department of Psychiatry, University of Illinois at Chicago, Chicago, IL

Abstract

Medical subject headings (MeSH) are a flexible and useful tool for describing biomedical concepts. Here, we present MeSHier, a tool for assigning MeSH terms to biomedical documents based on abstract similarity and references to MEDLINE records. When applied to PubMedCentral papers, NIH grants, and USPTO patents we find that these two sources of information produce largely disjoint sets of related MEDLINE records, albeit with some overlap in MeSH. When combined they provide an enriched topical annotation that would not have been possible with either alone. MeSHier is available as a demo tool that can take as input IDs of PubMed papers, USPTO patents, and NIH grants: http://abel.lis.illinois.edu/cgi-bin/meshier/search.py

Keywords

Controlled vocabularies; Medical subject headings

1. INTRODUCTION

Medical subject headings (MeSH) are a controlled vocabulary used by the National Library of Medicine for indexing documents in MEDLINE[16]. MeSH is comprised of terms arranged in a hierarchical structure covering a wide variety of topics[16]. It is important to note that the MeSH vocabulary is not limited to the biomedical domain; there are also terms covering topics in social science, the humanities, information science and beyond. The usefulness of MeSH continues to inspire a number of efforts to automatically predict MeSH terms for unlabeled documents. These efforts typically focus on predicting terms for unlabeled articles in MEDLINE. Here, we develop a generalized method for predicting

Request permissions from permissions@acm.org.

[*] Corresponding author: vtorvik@illinois.edu.

MeSH terms that can apply to documents beyond MEDLINE. In our initial experiments, we have focused primarily on NIH grants, USPTO Patents and PubmedCentral research papers.

Our method is developed upon the hypothesis that documents retrieved by abstract similarity and references are complementary. Our preliminary experiments find that the sets of documents retrieved using these two methods are nearly disjoint. This complementarity provides two distinct sources of evidence to use in predicting appropriate MeSH terms. Multiple factors contribute to the lack of overlap between these document sets. The leading factor is likely temporal; references are inherently bound by the publication date of a document. Therefore, retrieval by abstract similarity covers a larger temporal range and often retrieves different documents than by direct reference. Document sets can become further disjoint in certain domains. Here, we find that NIH grants often have fewer available references as they describe prospective work rather than past research. Our method necessarily makes greater use of abstract similarity rather than citations in grants. In patents, fundamental differences in vocabulary also likely lead to different documents being retrieved than through direct citation.

In addition to demonstrating a classification method, we present several applications that utilize MeSH in novel ways. An initial example: NIH grants are an important source of information about the structure and priorities of publicly funded science. These grants are useful for describing the prospective goals of research scientists, as well as a retrospective tool to compare research published under a grant to the original objectives of the grant. Though a great deal of information is available about biomedical grants, it is difficult to aggregate their content beyond broad descriptors. The hierarchical structure of MeSH terms permit a more granular description of what topics receive funding than is currently available. In the realm of patents, MeSH can play an important role in improving information retrieval. A key challenge for information retrieval in patents is overcoming the ambiguity and range of terms used for the same or similar concepts.

Finally, we test our methodology in a number of early experiments across several domains. With respect to NIH grants, we utilize categorical spending reports as a control. Since 2008, the Research, Condition, and Disease Categorization (RCDC) system has been used to match research projects to a set of descriptive categories. These categories typically reflect a disease or a research area (examples include ’cancer’ and ’biodefense’)[1]. We match each category to a MeSH term, and then assess the accuracy of our method. Additionally, we analyze a number of MEDLINE papers by closely comparing the predicted terms against the terms selected by indexers at the NLM. We also examine several patents, where our approach correctly identifies terms that do not appear directly in the text. We further compare the performance of our system against the NLM’s NLP-driven MeSH on Demand tool.

2. BACKGROUND

The MeSH vocabulary contains approximately twenty two thousand descriptors, organized in an eleven level hierarchy across sixteen categories [16]. Examples of these categories include anatomic terms, drugs and diseases and organisms. MeSH terms are assigned by expert indexers at the NLM. The annotation process uses a recommender system that provides indexers with suggestions that are then manually filtered. Most MEDLINE records have an average of 13 annotations per document, although this can vary depending on the domain [12].

The high cost of manually indexing MEDLINE records inspires a continued research interest in automatic classification methods. Numerous research groups have proposed MeSH prediction systems. K-nearest neighbor methods are among the most common approach [12, 24, 19, 13]. These methods typically only define neighbors with respect to citations. Another common approach leverages machine learning techniques to identify patterns between the document and MeSH terms [26, 25, 21]. One of the most popular approaches is to use natural language processing to extract MeSH terms from the text of the document directly. NLP methods have demonstrated some promise, but continue to be limited by the inherent difficulty of processing ambiguous text. The NLM’s MetaMap is one of the leading examples of the NLP approach. MetaMap extracts UMLS concepts, including MeSH, directly from text. More recent efforts have shown promise in this area by focusing on deep semantic representation techniques [18].

Perhaps the most well known MeSH prediction tool is the Medical Text Indexer (MTI) system. The MTI system assists NLM indexers in providing MeSH terms[17]. The MTI system takes inputs of an identifier, title and abstract but is also capable of processing arbitrary biomedical text [17]. Recommendations are computed using two methods: MetaMap indexing and a K-nearest neighbors algorithm that identifies similar citations [16]. MetaMap processes the title and abstract to identify UMLS Metathesaurus concepts that can then be mapped to MeSH. Precision and recall performance for the MTI system is typically around .60 [16].

Relatively few efforts have focused on predicting MeSH terms for documents outside of MEDLINE. Previous research teams have attempted to apply the MeSH vocabulary to patents. In ”Annotating Patents with Medline MeSH Codes via Citation Mapping” Thomas Griffin et. al presented a system which matched patent references to MED-LINE records and extracted MeSH terms [10]. This system retrieves the MeSH terms and arranges them alphabetically or by frequency of the term. A patent held by IBM titled ”System and Method for Annotating Patents with MeSH Data” proposes a similar procedure that extracts non-patent references directly using the MeSH vocabulary of the cited documents [3].

These approaches were both inspired by the clear information retrieval value of the MeSH vocabulary. Thomas Grin et. al [8] developed an analysis comparing the IPC classification system and MeSH, finding that the MeSH vocabulary is better suited to describing biomedical research. However, neither classification system discussed above attempts to rank or filter MeSH terms beyond frequency measures. Our system weights terms by their relative frequency within MEDLINE as well as in related papers in order to identify a smaller set of terms that are likely to be salient.

3. METHODOLOGY

Our approach leverages two pieces of complementary information from a document: its citations to the biomedical literature, and its concatenated title and abstract. The records cited by a document provide strong evidence to its own content because each reference reflects a conscious decision on the part of its authors. References formally signify a relationship from one work to another. Cited records clearly have limitations as a signal of a document's content: often the cited work may only have a limited or tangential relationship to the document. Depending on the domain, a document may have very few or no citations at all. Indeed, bibliometric analyses have shown that highly productive researchers tend to use fewer ”foundational” citations, and draw upon a broader body of knowledge[15]. A strictly citation driven methodology would be prone to systematic error in these cases. To mitigate these limitations, we also use the title and abstract text to identify related documents by text similarity. This allows us to identify those terms that are in ”agreement” between the two sources of information. While citations are inherently limited to work published in the past, text similarity allows to leverage documents published at any time. Retrieval by text similarity can partially compensate for cases where relevant literature was not cited in the original document. We observe a high degree of complementarity across the domains we have examined thus far.

Our approach is based on a suite of tools for locating MEDLINE documents from either cited records or from text. We use a tool based on our previous research called Absim to retrieve MEDLINE records based on BM25 similarity utilizing title and abstracts [23]. We retrieve cited records either directly (when available) or using Patci, a tool for matching citation strings to MEDLINE records [2]. The system determines which set of citation matching tools to use based on the input ID. For USPTO records, we use a collection of patents preprocessed using Patci to retrieve cited records.

Procedurally, our tool first identifies direct citations to MEDLINE records and then further collects the references of cited documents. MeSH terms are extracted and collected, with ”citations of citations” terms ranked lower than direct citations. Additionally, the tool collects the top 20 MEDLINE records by text similarity. The MeSH terms are extracted from these records as well, and weighted some-what lower than those from cited records. We collect further references from the abstract similarity set. Our tool then positively weights terms that appear in both the cited record and the abstract similarity set, to reflect ”agreement” between the two sources of data. We rank MeSH terms using an additive weighting function based on the frequency and distinctiveness of the MeSH terms. The system then returns the top 15 terms to the user, ranked by the following score:

$$score = M_{log\ frq}\beta_1 \mid Ack \mid + M_{log\ frq}\beta_2 \mid Ack_r \mid \quad (1)$$

$$+M_{log\ frq}\beta_3 \mid Abs \mid + M_{log\ frq}\beta_4 \mid Abs_r \mid \quad (2)$$

where M_{logfrq} is the log frequency of the MeSH term in MEDLINE. $|Ack|$ and $|Ack_r|$ correspond to the number of cited records and the number of cited records retrieved from the citation set (the ”citations of citations”). Likewise, $|Abs|$ and $|Abs_r|$ are the count of similar abstracts containing the term, and the term frequency count in the records cited by the similar abstract set. The β terms are heuristically selected weights. In the experiments detailed here, the weight of cited records and ”citations of citations” are 40 and 4, and the weight of similar abstracts is 20. The weight for cited records from similar abstracts is 1.

4. RESULTS AND EVALUATION

We conducted several experiments to test the performance of our methodology in different domains outside of MED-LINE, including NIH grants and USPTO patents. Additionally, we examined a number of MEDLINE papers and compared our system’s predicted terms against the terms selected by NLM indexers. Although our objective is to develop a classification system for documents outside of MED-LINE, predictions within MEDLINE are useful because there is a direct point of comparison between terms selected by manual annotators and our system.

A significant challenge in evaluating any MeSH prediction system is that there are relatively few sources of data to validate predictions. Even when term assignments do exist, it can be difficult to assess accuracy in a comprehensive manner. Assigning controlled vocabulary terms is an inherently subjective process that is shaped by normative practices among indexers. The complexity of the MeSH vocabulary complicates accuracy metrics because there may be several related terms that are appropriate matches, at varying levels of specificity. We used a variety of strategies to address these challenges, detailed in each section below.

4.1 PubMedCentral Quantitative Assessment

Although our focus here is on non-MEDLINE records, we conducted several experiments to evaluate the performance of our method on MEDLINE records annotated with MeSH. We collected a data set comprising 1600 papers published between 2000 and 2015. We retrieved every paper with an abstract and at least one citation and at least one MeSH term for each year. From this set, we randomly selected 100 papers. We processed each paper with MeSHier, recording the log cumulative frequency within MEDLINE and the frequency counts for each citation category (the number of times the term appeared in acknowledged references, references of references, similar documents, and acknowledged references of similar documents). We also included any terms assigned to the paper not captured by our retrieval process. Due to changes in the MeSH vocabulary over time, there are some terms where we did not have frequency information available. For these, we assigned a log frequency of 0.3.

We subsequently trained a logistic regression model on several subsets of features: using only acknowledged references, only similar documents, and both together. We then evaluated the accuracy of the top 15 scoring terms for each model. The performance and recall characteristics of these models are detailed in Table 2. We observed that predicted terms that were not direct matches were often conceptually similar to assigned term, or otherwise relevant to the paper, indicating that the precision and recall measures are likely to be a lower bound estimate of classifier performance. Due to the complexity in assessing performance, we also include a case study of several individual papers.

4.2 PubMedCentral Case Studies

Closer comparison between predicted MeSH terms and actual MeSH terms in MEDLINE itself is instructive. Table 3 collects four papers, spanning topics in informatics and genomics. For each paper, we have compared our predictions against the MeSH terms assigned by indexers. These papers were selected from our past publications in order to ensure that we can accurately assess whether or not the predicted terms are appropriate for the paper. Expert manual assessment of the predicted terms is necessary because in many cases the predicted terms are actually more descriptive than those assigned manually.

An example of this can be found in the first two papers listed in the table. Both of these are informatics papers related to author name disambiguation. Both papers have relatively few assigned MeSH terms. Our system highly ranks terms like ’Authorship’, ’Publishing’, ’Bibliometrics’ and ’Cooperative Behavior’. These terms are more descriptive of the content of these papers than terms like ’Algorithms’ or ’Names’. There is only one instance of a wholly inaccurate prediction. The lowest ranked term ’Nerve net’ for the first paper refers to biological networks. This term is likely introduced by abstract similarity, as the paper discusses networks in the sense of mathematical graph theory, rather than biological neural nets.

The second set of papers are genomics and bioinformatics papers. There is a high degree of overlap between predicted terms and actual terms. Notably, the predicted terms for ”Natural antisense transcripts are co-expressed with sense mRNAs in synaptoneurosomes of adult mouse forebrain” includes Alzheimer disease as the leading term. Indeed, the abstract of the paper contains the sentence ”Several of these pairs involve mRNAs that have been implicated in synaptic functions and in Alzheimer disease pathways.” Alzheimer disease does not appear in the list of MeSH terms selected by indexers. In this case, the predicted terms identify an important topic that is excluded in the manual annotation. A broader pattern appears to be that our system tends to emphasize subjects (diseases, processes and phenomena) and somewhat discounts methods. This pattern is similar to what was observed in the NIH grant evaluation. This tendency may be explained by a relatively larger degree of ”agreement”in subjects, and a more divergent set of methods in similar papers. In other words, contrasting methodological approaches may lead to a lower ranking against a more tightly focused set of terms on subjects.

Through this small set of case studies, it seems that further work is required to elucidate the performance of the system in a broader context. Informatics is a relatively minority within MEDLINE; it may be the case that MeSH terms are generally sparser and less descriptive in these fields.

An important result appears to be that predicted terms may sometimes be more descriptive than indexed terms. This indicates that caution should be applied in evaluating classifier performance strictly on how well the system replicates human annotation; there may be multiple valid classifications for any given document and it may be shortsighted to consider alternative annotations merely as ”errors.”

4.3 NIH Grants

We used the NIH Categorical Spending reports to evaluate our prediction system for grant records. We collected a set of 21 grants, covering every NIH institute. The following application numbers were used: 7888075-AG, 7938182-AI, 7861111-AR, 7847187-AT, 8055735-DC, 8013667-DE, 7865069-DK, 8015869-EB, 7948476-ES, 7905426-EY, 7919006-GM, 7984978-GM, 7949120-HD, 7948564-HG, 7937672-HL, 8073856-MD, 8004908-MH, 8023344-MH, 8073362-NR, 8068994-NS. Note that the institute code is indicated after each id with a dash. For each of these grants, we extracted the assigned funding categories and matched each term to a corresponding MeSH heading. We then processed the grant using MeSHier and compared the top 15 ranked MeSH terms with the assigned term in each category.

The NIH funding category terms are intended to broadly describe the content of a grant. Because controlled vocabulary terms are typically used to describe only the more distinctive aspects of a paper, we hypothesized that many of the predicted terms would be related to the funding category terms, but in many instances would be more detailed. We also anticipated that the more distinct funding category terms (e.g. the fairly specific ”ALS” vs the more general ”behavioral science”) would be more likely to directly match a predicted term.

Due to the varying levels of descriptive granularity we defined matches in five ways: no match, a direct match, a direct child match, a match to a closely related term, and a match to a more distantly related term. In many instances, several predicted terms could be considered a match. For each paper, we record the highest ranking match except in cases of a direct match. For records with a direct match, we record the rank of that term even if a higher ranking term matched indirectly.

A direct match meant that the predicted term appears in the list exactly. A direct child match applies when a term is predicted that is a direct child term in the MeSH hierarchy of the spending category term. A closely related match was used to indicate a term that is clearly related to the concept but is neither an exact match or a direct child. A distant match was any match that captured the broad area of the category term.

In our test sample, we matched 57.75% (67/116) of the category spending terms. The average rank of the matching term was 5.3. As discussed above, the type of match was weighted strongly towards ’closely related’ matches. Table 5 lists the frequency of match type. Exact and direct sibling matches made up approximately 11% of total matches. We propose that the relatively small number of exact matches is due to the generality of the funding category terms. In many cases, the closely related terms were related to more specific concepts; for example ’Gene regulatory networks’ is clearly a match to the category of ’genetics’, but is more narrowly defined.

Every grant in our test sample matched at least one funding category. The average match rate of assigned categories across papers was 64.89%. A common pattern we observed is that highly general categories (”Behavioral Science” and ”Biotechnology”) were not frequently matched in our predictions. We hypothesize that this is also related to the specificity differences discussed above; in some sense these terms are so broad that they would be ubiquitous if included by indexers.

The funding categories used in the sample matched to eight top-level MeSH categories. Not all top-level categories were matched at the same rate; the ”Phenomena and Processes”and ”Diseases”category were retrieved more frequently. The ”Technology and Agriculture” category was never retrieved, and ”Informatics” and ”Health Care” were perfectly retrieved but appeared in only a few grants. These were likely retrieved due to their relative distinctiveness in the biomedical literature. The failure to capture the ”Technology and Agriculture” grouping is likely due to the broadness of description problem described above. The most common term in this category was ”Biotechnology.”

4.4 USPTO Patents

Unlike NIH grants and MEDLINE papers, it is difficult to directly assess the validity of predicted terms in patents other than through close reading. For that reason, we include a case study of three biomedical patents, collected in Table 6. These patents were selected in connection with other projects. The first patent, US7262047 refers to an extremophile species of bacteria that is resistant to UV sterilization. The title and abstract are both contained at the top of the page. Table 6 contains both our system’s predicted terms, and terms predicted by the NLM’s MeSH on Demand tool.

It is difficult to determine from the abstract alone that the patent discusses a species of bacteria that has the potential to survive on spacecraft, introducing a potential risk of ’forward contamination’ in missions to Mars. Despite this lack of information in the abstract text, terms like ’spacecraft’, ’Mars’ and ’Extraterrestrial Environment’ are included in the list of predicted terms. These terms are largely retrieved through citations. This particular patent exemplifies the risk of using only text information. The NLP-driven MeSH on Demand tool misses these terms because they do not appear in the abstract. However, the combination of citation and text information together successfully recovers important information about the patent.

Conversely, NLP techniques can successfully identify content not easily retrievable using our methodology. In Patent US6136858, a patent related to infant formula, MeSH on Demand successfully retrieves concepts related to oils discussed in the abstract that our system misses. However, our system retrieves important concepts missed by the NLP technique, such as ”milk” and ”defecation.” In this case, the particular combination of oils constitute the invention; therefore the terms related to the oils likely do not appear frequently in the cited or similar literature.

Finally, US5719064 concerns a patent related to diagnostic techniques for treatment of ankylosing spondylitis. Our system identifies key concepts such as ”HLA-B27” and ”ankylosing spondylitis,” and ”klebsiella.” We note again a case where our system detects an important concept (”klebsiella”, a potentially causal agent involved in ankylosing spondylitis) where it is not directly mentioned in the abstract text. Our system also retrieves the more specific ”ankylosing spondylitis” vs the more generic ”spondylarthopathies”.

5. APPLICATIONS

5.1 Economic Analysis of Scientific Funding

Traditional methods of tracking the impact of National institutes of Health (NIH) funding have focused on examining the output of single grants or aggregating obligations by disease area. Although these methodologies are able to highlight important areas of research and the extent to which they receive federal support, they do not answer larger questions concerning the impact of NIH funding on entire fields of research and across fields. Addressing such questions requires that NIH grant awards be linked to some comprehensive and rigorous classification of fields, with MEDLINE’s Medical Subject Headings (MeSH) taxonomy being a natural choice. Elsewhere, we have constructed text and citation-based metrics that measure the transformativeness and impact of research, which can be used to assess the quantity and quality of output both within and across fields. These provide rich measures of the value of scientific research for understanding the overall impact of NIH funding on the nation’s science and innovation ecosystem.

NIH Associate Director Carrie D. Wolinetz, Ph.D. noted in 2016 that ”The pathways from research to practice to changes in public health are typically non-linear and unpredictable. For a scientific discovery to make that journey may take decades or more and involves a complex ecosystem”[27]. Wolinetz’s characterization of the research process emphasizes the importance of developing a broad-based methodology for tracking the flow of federal funding across disciplinary fields. The case of Neurostimulation Technologies, provides a particularly illustrative case study of how NIH funding can generate ideas that transcend traditional disciplinary boundaries and impact the overall science ecosystem[7]. Beginning in the late 1960s, researchers supported by NIH began experimenting with using electrodes for the purpose of restoring hearing loss which evolved into more advanced cochlear implants by the mid-1990s[7]. Motivated by the initial research auditory rehabilitation, researchers in 1973 began examining the relationship between electrical function and Parkinson’s disease eventually leading to the development of treatments that successfully reduced the intensity of tremors[7]. Recently, this same research has served as a foundation for new methods for treating spinal cord injuries and vision loss[7].

Connecting NIH funding with the MeSH taxonomy creates a foundation for a systematic examination of the impact of federal funding on the research and innovation ecosystem in terms of the generation and flow of both scholars and ideas. Theory suggests that researchers and scientists may respond to economic incentives, such as streams of research funding, by rationally adjusting their behavior. As a direct effect of federal funding, we expect the quantity and quality of work to increase in fields that are NIH target areas. As a secondary effect, we should also expect that researchers at the margin between fields may rationally redirect their activities in response to federal funding. As mentioned, the advantage of linking NIH grant abstracts to the MeSH taxonomy is that we are able to link these data with Medline and Thompson Reuter’s Web of Science (WoS). The text and citation-based metrics derived from these databases will allow us to examine the flow of researchers across fields in addition to the generation and dissemination of ideas across fields.

5.2 Information Retrieval in USPTO Patents

There are numerous challenges for effective information retrieval of biomedical patents[22, 8]. Past efforts have focused on creating resources that describe patents using semantic web techniques in order to create search systems that are better suited to biomedical researchers[22]. Patent information retrieval is inherently complex: patents are simultaneously a technical record and a legal document defining intellectual property rights. As a result, patent language can be difficult to understand and search without both technical and legal expertise. One major contributor to this complexity is that patents do not currently have a robust classification system with respect to biomedical concepts[8].

We are currently exploring information retrieval systems using automatically assigned MeSH terms to improve accessibility of patents. By allowing researchers to use MeSH, an already familiar vocabulary, we simplify the search process and avoid the necessity for complex queries covering many variants of a concept.

6. DISCUSSION

The classification method described above successfully predicts MeSH terms for a variety of biomedical documents outside of MEDLINE. The key advantage of this approach is that it is sufficiently general to be used in many different domains. Further work is required to identify if and how the performance of our technique varies systematically between domains. In our evaluation of our technique within NIH grants, we find that the predicted terms are often more specific than the funding categories we used as a control. Likewise, experiments on MEDLINE found that the predicted terms are sometimes more descriptive than the manually assigned terms. One consistent finding is that assessing accuracy is a complex endeavor in MeSH prediction. Further work is also required to assess the degree of complementarity in retrieved document sets across domains.

Acknowledgments

Research reported in this publication was supported in part by NIH National Institute on Aging P01AG039347. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. [Accessed: 2016-01-26] The research, condition, and disease categorization process. 2016. https://report.nih.gov/rcdc/index.aspx
2. Agarwal S, Lincoln M, Cai H, Torvik VI. [Accessed: 2016-01-26] Patci: A probabilistic citation matcher. http://abel.lis.illinois.edu/cgi-bin/patci/search.pl
3. I. B. M. Corporation. System and method for annotating patents with mesh data, 2007. US Patent Application. US 20070112833 A1.
4. DeLong M. Activity of pallidal neurons during movement. Journal of Neurophysiology. 1971; 34:414–427. [PubMed: 4997823]
5. DeLong M. Activity of basal ganglia neurons during movement. Brain Research. 1972; 40:127–135. [PubMed: 4624486]

6. DeLong M. Putamen: activity of single units during slow and rapid arm movements. Science. 1973; 179:1240–1242. [PubMed: 4631890]
7. Dodson S. Nih case studies of research impact. Science of Science Policy. 2016
8. Eisinger D, Tsatsaronis G, Bundschus M, Wieneke U, Schroeder M. Automated patent categorization and guided patent search using ipc as inspired by mesh and pubmed. Journal of Biomedical Semantics. 2013; 4(1):S3. [PubMed: 23734562]
9. Gerasimenko YP, Lu DC, Modaber M, Zdunowski S, Gad P, Sayenko DG, Morikawa E, Haakana P, Ferguson AR, Roy RR, et al. Noninvasive reactivation of motor descending control after paralysis. Journal of neurotrauma. 2015; 32(24):1968–1980. [PubMed: 26077679]
10. Griffin TD, Boyer SK, Councill IG. Annotating Patents with Medline MeSH Codes via Citation Mapping. Springer New York; New York, NY: 2010. 737–744.
11. House UJWF. Long term results of electrode implantation and electronic stimulation of the cochlea in man. The Annals of otology, rhinology, and laryngology. 1973; 82:504–517.
12. Huang M, Névéol A, Lu Z. Recommending mesh terms for annotating biomedical articles. Journal of the American Medical Informatics Association. 2011; 18(5):660–667. [PubMed: 21613640]
13. Kim W, Aronson RA, Wilbur JW. Proceedings of the AMIA Symposium. American Medical Informatics Association; 2001. Automatic mesh term assignment and quality assessment; 319
14. Lin J, Wilbur W. Pubmed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics. 2007; 8
15. Milojević S. How are academic age, productivity and collaboration related to citing behavior of researchers? PloS one. 2012; 7(11):e49176. [PubMed: 23145111]
16. Mork JG, Demner-Fushman D, Schmidt S, Aronson AR. Recent enhancements to the nlm medical text indexer. Working Notes for CLEF 2014 Conference; Shefield, UK. 2014. 1328–1336.
17. Mork JG, Jimeno-Yepes A, Aronson AR. The nlm medical text indexer system for indexing biomedical literature. BioASQ@ CLEF. 2013
18. Peng S, You R, Wang H, Zhai C, Mamitsuka H, Zhu S. Deepmesh: deep semantic representation for improving large-scale mesh indexing. Bioinformatics. 2016; 32(12):i70. [PubMed: 27307646]
19. Ruch P. Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics. 2006; 22(6):658–664. [PubMed: 16287934]
20. Simmons FB, Mongeon CJ, Lewis WR, Huntington DA. Electrical stimulation of acoustical nerve and inferior colliculus: Results in man. Archives of OtolaryngologyâĂŤHead & Neck Surgery. 1964; 79(6):559.
21. Sohn S, Kim W, Comeau DC, Wilbur WJ. Optimal training sets for bayesian prediction of mesh® assignment. Journal of the American Medical Informatics Association. 2008; 15(4):546–553. [PubMed: 18436913]
22. Sougata Mukherjea BB. Biopatentminer: An information retrieval system for biomedical patents. Proceedings of the 30th VLDB Conference; 2004.
23. Torvik VI. [Accessed: 2016-01-26] Absim: A tool for calculating bm25 similarity among pairs of abstracts in pubmed. http://abel.lis.illinois.edu/cgi-bin/absim/search.py
24. Trieschnigg D, Pezik P, Lee V, De Jong F, Kraaij W, Rebholz-Schuhmann D. Mesh up: effective mesh text classification for improved document retrieval. Bioinformatics. 2009; 25(11):1412–1418. [PubMed: 19376821]
25. Wahle M, Widdows D, Herskovic JR, Bernstam EV, Cohen T. Deterministic binary vectors for efficient automated indexing of medline/pubmed abstracts. AMIA annual symposium proceedings; American Medical Informatics Association; 2012. 940
26. Wilbur WJ, Kim W. Stochastic gradient descent and the prediction of mesh for pubmed records. AMIA Annual Symposium Proceedings; American Medical Informatics Association; 2014. 1198
27. Wolinetz C. Capturing impact: A method for measuring progress. 2016

Table 1

NIH grant funding category label accuracy, aggregated to the top MeSH category

MeSH Top Level Category	**Accuracy**
Phenomena and Processes	0.71
Diseases	0.57
Technology, Industry, Agriculture	0.00
Psychiatry and Psychology	0.56
Health Care	1.00
Information Science	1.00
Anthropology, Education, Sociology	0.25
Disciplines and Occupations	0.80
Analytical, Diag., Therapeutic Tech.	0.46

Table 2

Comparison of Model Performance in Pub-MedCentral

Model	Precision	Recall	F_1 Score
Citation Only	0.41	0.47	0.44
Similarity Only	0.39	0.45	0.42
Combined	0.43	0.50	0.46

Table 3

Comparison of Predicted MeSH terms for Selected Papers in PubmedCentral

PMID	Title	Predicted MeSH	Actual MeSH
23894639	Has large-scale named-entity network analysis been resting on a flawed assumption?	Authorship; Patents as Topic; Bibliometrics; Publishing; Models, Theoretical; MEDLINE; **Algorithms; Names**; Cooperative Behavior; Research; Periodicals as Topic; Neural Networks (Computer); Computer Simulation; Research Personnel; Nerve Net	**Algorithms; Names**; Publications
14728536	A probabilistic similarity metric for Medline records: a model for author name disambiguation	Periodicals as Topic; **MEDLINE; Medical Subject Headings**; Abstracting and Indexing as Topic; Publishing; **Authorship; Bibliometrics**; Information Storage and Retrieval; United States; Subject Headings; Randomized Controlled Trials as Topic; Semantics; Databases, Bibliographic; Algorithms; Internet	**Authorship; Bibliometrics; MEDLINE; Medical Subject Headings**; Names; Probability
15453917	A population-based statistical approach identifies parameters characteristic of human microRNA-mRNA interactions	**MicroRNAs**; Base Sequence; **RNA, Messenger**; Molecular Sequence Data; RNA, Small Interfering; Nucleic Acid Conformation; Caenorhabditis elegans; RNA Interference; Gene Silencing; 3' Untranslated Regions; Gene Expression Regulation, Developmental; RNA, Double-Stranded; Sequence Homology, Nucleic Acid; RNA, Untranslated; **Computational Biology**	**Computational Biology**; Genetics, Population; Humans; **MicroRNAs; RNA, Messenger**
18812194	Natural antisense transcripts are co-expressed with sense mRNAs in synaptoneurosomes of adult mouse forebrain	Alzheimer Disease; **Synapses; Mice; RNA, Messenger**; Transcription, Genetic; Neurons; Dendrites; **Molecular Sequence Data**; RNA, Antisense; MicroRNAs; Hippocampus; Rats; Brain; Amyloid beta-Peptides; **Amyloid Precursor Protein Secretases**	**Amyloid Precursor Protein Secretases** ; Animals; Aspartic Acid Endopeptidases; Humans; Male; **Mice** ; Mice, Inbred C57BL; **Molecular Sequence Data**; Prosencephalon; **RNA, Messenger**; RNA, Untranslated; Sirtuins; Subcellular Fractions; **Synapses**; Synaptosomes; rab GTP-Binding Proteins;

Table 4

MeSH terms that differ from funding category label. Note that 24 exact matches between the funding label and MeSH terms are omitted

Funding Category Label	Selected MeSH Term
Nutrition	Diet
Dental/Oral and Craniofacial Disease	Stomatognathic Disease
Human Genome	Genome
Networking and Information Technology	Informatics
Complementary and Alternative Medicine	Complementary Therapies
Diagnostic Radiology	Radiography
Breast Cancer	Breast Neoplasms
Child Abuse and Neglect Research	Child Abuse
Emerging Infectious Diseases	Communicable Diseases
Biodefense	Biological Warfare Agents
Injury Accidents/Adverse Effects	Wounds and Injuries
Comparative Effectiveness	Comparative Effectiveness Research
Autism	Autistic Disorder
Behavioral and Social Science	Behavioral Sciences
Orphan Drug	Orphan Drug Production
Basic Behavioral and Social Science	Behavioral Sciences
Violence Against Women	Domestic Violence
Infectious Diseases	Communicable Diseases
Diabetes	Glucose Metabolism Disorders
Digestive Diseases	Digestive System Diseases
Chronic Liver Disease and Cirrhosis	Liver Cirrhosis
Clinical Trials	Clinical Trials as Topic
Eye Diseases and Disorders of Vision	Eye Diseases
Vector-borne diseases	Communicable Disease Control
Coronary Heart Disease	Coronary Disease
Cancer	Neoplasms
Pain Conditions - Chronic	Chronic Pain
Alzheimer’s Disease	Alzheimer Disease
Substance Abuse	Substance-Related Disorders
Brain Disorders	Brain Diseases
Clinical Research	Clinical Protocols
Mental Retardation	Intellectual Disability
ALS	Amyotrophic Lateral Sclerosis
Cardiovascular	Cardiology

Table 5

Frequency of Match Type

Match Category	Proportion
No Match	0.42
Exact Match	0.08
Direct Sibling	0.04
Closely Related	0.41
Distantly Related	0.05

Table 6

MeSHier vs Mesh On Demand predictions

Patent ID	Predicted Terms	MeSH on Demand Predictions
US7262047	Spores, Bacterial; Bacillus subtilis; **Bacillus**; Space-craft; Extraterrestrial Environment; Ultraviolet Rays; DNA, Bacterial; Mars; Bacterial Proteins; Spectrometry, Mass, Matrix-Assisted Laser Desorption-Ionization; Phylogeny; **DNA, Ribosomal**; Molecular Sequence Data; RNA, Ribosomal, 16S; Bacillus cereus	**Bacillus; DNA, Ribosomal**; Databases, Nucleic Acid; Inventions; Nucleic Acid Hybridization; Sequence Analysis; Sterilization ; Sterilization, Reproductive
US6136858	Infant Food; Milk, Human; Infant; Dietary Fats; Breast Feeding; **Infant**, Newborn; Feces; Milk; Antibodies, Bacterial; Nucleotides; Humans; Food, Formulated; **Fatty Acids**; Defecation; Nitrogen	Carbohydrates; **Fatty Acids**; Humans; **Infant**; Infant Formula; Inventions; Minerals; Plant Oils; Safflower Oil; Soybean Oil; Vitamins
US5719064	**HLA-B27 Antigen**; Spondylitis, Ankylosing; Amino Acid Sequence; Molecular Sequence Data; Epitopes; HLA Antigens; Cross Reactions; Klebsiella pneumoniae; Antibodies, Bacterial; Arthritis, Reactive; Antigens, Bacterial; **T-Lymphocytes**; beta 2-Microglobulin; Base Sequence; Mice	Alleles; Amino Acid Motifs; Amino Acid Sequence; Amino Acids; Enterobacteriaceae; Gram-Negative Bacteria; **HLA-B27 Antigen**; Peptides; Spondylarthropathies; **T-Lymphocytes**