

Topic Models to Interpret MeSH – MEDLINE’s Medical Subject Headings

David Newman^{1,2}, Sarvnaz Karimi¹, and Lawrence Cavendon¹

¹ NICTA and The University of Melbourne, Victoria, Australia

² University of California, Irvine, USA

{david.newman,sarvnaz.karimi,lawrence.cavendon}@nicta.com.au

Abstract. We show how topic models are useful for interpreting and understanding MeSH, the Medical Subject Headings applied to articles in MEDLINE. We show how our resampled author model captures some of the advantages of both the topic model and the author-topic model. We demonstrate how the topic modeling approach can provide an alternative and complementary view of the relationship between MeSH headings that could be informative and helpful for people searching MEDLINE.

1 Introduction

MeSH are the subject headings used for tagging articles in MEDLINE, the largest biomedical literature database in the world. PubMed – the interface for searching MEDLINE – extensively uses these MeSH headings. Most PubMed queries are mapped to queries that involve MeSH headings, e.g. the query “teen drug use” gets mapped to a longer query that searches for the MeSH headings “Adolescent” and “Substance-Related Disorders” (this mapping is explained in [1]). Therefore, it is critical for researchers and health-care professionals using PubMed to understand what is meant by these MeSH headings, since MeSH headings have a direct effect on search results.

One may try to understand MeSH headings by understanding how MeSH headings are applied to articles. However, MeSH tagging is a complex procedure performed by a team of expert catalogers at the National Library of Medicine in the US³. These catalogers use a system of tools including NLM’s MetaMap, which maps biomedical text to the UMLS Metathesaurus, and other machine learning tools that score and suggest MeSH headings.

We take a statistical approach to this problem, using topic modeling of large sets of PubMed search results to provide a semantic interpretation of MeSH headings. Through large scale patterns of MeSH tagging, and patterns of co-occurring words in titles and abstracts, we independently learn the meaning of MeSH terms in a data-driven way. While this paper focuses on MEDLINE and MeSH, this framework is more broadly useful for any collection of text documents that is tagged with subject headings.

³ MeSH tagging is described in detail at <http://ii.nlm.nih.gov/mti.shtml>

MeSH heading	Major MeSH heading	Major qualifier	MeSH-qualifier combination
Humans	Brain	metabolism	Signal Transduction (physiology)
Female	Breast Neoplasms	physiology	Antineoplastic Combined Chemotherapy Protocols (therapeutic use)
Male	Neoplasms	genetics	Magnetic Resonance Imaging (methods)
Animals	Apoptosis	methods	Apoptosis (drug effects)
Adult	HIV Infections	chemistry	Neurons (physiology)
Middle Aged	Neurons	pharmacology	DNA-Binding Proteins (metabolism)
Aged	Signal Transduction	therapeutic use	Transcription Factors (metabolism)
Adolescent	Antineoplastic Agents	pathology	Antineoplastic Agents (therapeutic use)
Mice	Magnetic Resonance Imaging	immunology	Anti-Bacterial Agents (pharmacology)
Child	Anti-Bacterial Agents	diagnosis	Brain (metabolism)

Table 1. Most frequent MeSH headings, major MeSH headings, major qualifiers and MeSH-qualifier combinations in articles published since 2000.

Background on MeSH headings: MeSH headings are arranged in a large, complex and continually evolving hierarchy. Currently there are over 25,000 MeSH terms arranged in a DAG which includes a root and 11 levels. On average there are 16 MeSH headings attached to a MEDLINE article. All MeSH tags on a given article have an additional attribute MajorTopicYN which can take on the value *Y* or *N*, indicating whether the MeSH tag is the primary focus of the article. Furthermore, each application of a MeSH tag on an article may be qualified using zero, one, or more qualifiers, e.g. one could qualify the MeSH tag *Methadone* with the qualifier *therapeutic use*. There are over 80 qualifiers, but only a specific subset of qualifiers may be used with each MeSH heading. Qualifiers applied to articles also always have the attribute MajorTopicYN.

To gain some familiarity with the usage of MeSH headings and qualifiers, we provide lists of most frequent terms in Table 1. Rows in the table do not correspond – the four columns are separate. The first column shows the most frequent MeSH headings, irrespective of MajorTopicYN. We see headings that act as “check tags” (e.g. Human), used to restrict search results to certain classes of interest. The second column shows the most common *major* MeSH headings, where the heading or one of its qualifiers has MajorTopicYN=*Y*. Here we see a broad range of topics, covering both conditions/diseases (Neoplasms, HIV) and basic research (Neurons, Apoptosis, Signal Transduction). The most frequent qualifiers also span a wide range, and finally we see the top MeSH heading-qualifier combinations partly overlap with the most frequent major MeSH terms, providing more detail about what is prevalent in the published literature. For the rest of this paper, we only consider major MeSH headings (on average, 5 out of 16 MeSH headings applied to an article are major, or have a major qualifier).

2 Interpreting MeSH headings with topic models

2.1 Methods and data

Topics – learned by topic models – provide a natural basis for representing and understanding MeSH headings. Topic models (also known as Latent Dirichlet Allocation models or Discrete PCA models) are a class of Bayesian graphical

Label	PubMed query	# results
burns	Burns[MeSH Terms] AND Humans[MeSH Terms]	19,970
dopamine	Dopamine[MeSH Terms]	33,223
drug	Substance-Related Disorders[MeSH Terms] AND Adolescent[MeSH Terms]	22,361
imaging	Imaging, Three-Dimensional[MeSH Terms]	21,858
p53	p53[All Fields]	47,327
smoking	Smoking[MeSH Terms]	63,101

Table 2. PubMed queries run to produce query results sets for experiments. The number of results shown only count search results that contain abstracts.

models for text document collections represented by bag-of-words (see [2–4]). In the standard topic model, each document in the collection of D documents is modeled as a multinomial distribution over T topics, where each topic is a multinomial distributions over W words, and both sets of multinomials are sampled from a Dirichlet.

Rather than learn a single topic model of all of MEDLINE (an impractical task, especially given that we would need to learn thousands of topics), we chose to demonstrate our methodology using six query results sets shown in Table 2. We created PubMed queries that returned a large number of articles (10,000 to 100,000 search results) in a broad area, thus allowing us to get a large sample of MeSH tags used in that area. For topic modeling purposes, we only used PubMed search results that contained abstracts (many pre-1980 MEDLINE citations do not contain abstracts).

2.2 Topic model and author-topic model

We start with two topic models appropriate for our task: the standard topic model, and the author-topic model ([3, 5]). In the author-topic model, we are using MeSH headings as authors of the documents (using the obvious analogy that like an author, the MeSH heading is responsible for generating words in the title and abstract). To learn the model parameters we use Gibbs sampling. The Gibbs sampling equations for the topic model and author-topic model are given by

$$p(z_{id} = t | x_{id} = w, \mathbf{z}^{-id}) \propto \frac{N_{wt}^{-id} + \beta}{\sum_w N_{wt}^{-id} + W\beta} \frac{N_{td}^{-id} + \alpha}{\sum_t N_{td}^{-id} + T\alpha}, \quad (1)$$

$$p(z_{id} = t, y_{id} = m, | x_{id} = w, \mathbf{z}^{-id}, \mathbf{y}^{-id}) \propto \frac{N_{wt}^{-id} + \beta}{\sum_w N_{wt}^{-id} + W\beta} \frac{N_{tm}^{-id} + \gamma}{\sum_t N_{tm}^{-id} + T\gamma}, \quad (2)$$

where $z_{id} = t$ and $y_{id} = m$ are the assignments of the i^{th} word in document d to topic t and author m respectively, and $x_{id} = w$ indicates that the current observed word is word w . \mathbf{z}^{-id} and \mathbf{y}^{-id} are the vectors of all topic and author assignments not including the current word, N_{wt} , N_{td} and N_{tm} represent integer count arrays (with the subscripts denoting what is counted), and α , β and γ are Dirichlet priors. From the count arrays, we estimate the conditional distributions using

$$p(w|t) = \frac{N_{wt} + \beta}{\sum_w N_{wt} + W\beta}, p(t|d) = \frac{N_{td} + \alpha}{\sum_t N_{td} + T\alpha}, p(t|m) = \frac{N_{tm} + \gamma}{\sum_t N_{tm} + T\gamma}. \quad (3)$$

We use a MeSH heading’s distribution over topics, $p(t|m)$, as the canonical way to represent a MeSH heading using learned topics. The author-topic model directly estimates this distribution over topics for each MeSH heading. In the topic model, we estimate $p(t|m)$ by summing over the documents using $p(t|m) = \sum_d p(t|d)p(d|m)$, where $p(d|m)$ is the empirical distribution of observed application of MeSH headings to documents.

For each of the six query results sets, we learned topic and author-topic models and computed $p(t|m)$ for all the major MeSH headings that occurred in at least 10 articles in the query results set. The following examples show distributions over topics for three MeSH headings (with query set indicated with ‘q=’):

Alcohol-Related-Disorders [q=drug]
(0.27) [t26] drinking alcohol alcohol-use problem alcohol-related drinker women alcohol-consumption heavy ...
(0.11) [t36] dependence use-disorder criteria dsm-iv symptom diagnostic interview treatment adolescent ...

Artificial-Intelligence [q=imaging]
(0.39) [t62] segmentation shape feature classification detection structure automatic analysis representation ...
(0.32) [t71] image algorithm model object proposed framework approach problem propose estimation ...
(0.17) [t110] approach surface application efficient problem demonstrate texture component computation ...

Tobacco-Smoke-Pollution [q=smoking]
(0.31) [t98] passive air tobacco-smoke ets pollution environmental exposure active home indoor ...
(0.10) [t129] smoking smoker tobacco smoke consumption tobacco-use daily smoked cigarette current ...
(0.09) [t70] exposure exposed effect level environmental chemical relationship evidence observed dose ...
(0.08) [t120] policies policy ban smoke-free workplace law public restaurant restriction smoking ...

Under each MeSH heading we list the most probable topics according to $p(t|m)$. We denote a topic by a topic ID (e.g. [t26]) which has no external meaning, then the list of most likely words in that topic, followed by an ellipsis to indicate that the cutoff for printing words is arbitrary. The number preceding the topic ID is $p(t|m)$ for that topic. Topics accounting for less than 0.05 probability mass are not shown. This example shows that we learn sensible distributions over topics for these three MeSH headings. Note that the topics learned, and the association of topics with MeSH headings, is not completely independent of the results set returned by the query. For example, the topics associated with the MeSH heading Artificial-Intelligence are clearly oriented towards imaging.

When topic modeling, we want to learn “essential” topics, i.e. topics that are robust (albeit latent) features present in the data, which are reliably and repeatably found by a variety of techniques. However, even with the two closely-related models, the topic model and the author-topic model, we learn topics that are close, but clearly different. For example, for the burns query results set, we learn the following topics relating to children:

(1.1%) [t11] children pediatric year child age **burned** parent month **young** childhood **adult** infant mother **burn** ...
(0.7%) [t25] children child **abuse** parent pediatric **scald** mother year age **physical home** month infant childhood ...

where [t11] is learned by the topic model and [t25] by the author-topic model. While not shown, these topics are highly repeatable over different random initializations. The gist of these topics is clearly different (the different words are

bold), with the author-topic model learning an abuse variation to the topic. There is also a difference in the prevalence of the two topics, with the first topic accounting for 1.1% of all words, and the second topic accounting for 0.7% of all words. So one may be left wondering which is the better or more correct topic.

In practice, different topic models produce different topics and different statistics, which may not be obvious from the model formulations, but may be revealed by experiments. Figure 1 shows that the distribution of topic sizes for the topic model is flatter than that from the author-topic model for our data sets.

2.3 Resampled author model

There may be several reasons for preferring topics learned by the standard topic model. One could argue that the simpler model learns topics that are in some way more fundamental to the collection. Furthermore, even with our MEDLINE abstracts, we have ambiguity over authors: Are they the MeSH headings, or actual authors of the articles? We also may prefer the flatter distribution of topic sizes for better division and faceted searching of the collection.

Here we introduce the resampled author model. The resampled author model is the author-topic model run with a fixed word-topic distribution previously learned by the topic model:

$$p(y_{id} = m | x_{id} = w, z_{id} = t, \mathbf{z}^{-id}, \mathbf{y}^{-id}) \propto p(w|t) \frac{N_{tm}^{-id} + \gamma}{\sum_t N_{tm}^{-id} + T\gamma} \quad (4)$$

with $p(w|t)$ given by (3). The idea behind the resampled author model is to keep and use topics learned by the topic model, but learn a better association between topics and authors (in our case MeSH headings), than the naive computation of summing over documents. Indeed, our experimental results shown in Figure 2 show that the resampled author model does produce results that combines the learned topics from the topic model, and the relatively low entropy of topic distributions computed by the author-topic model.

3 Analysis of MeSH headings

Our topic representation of MeSH headings is useful for a broad array of tasks. First they are a direct way of explaining or interpreting what is meant by a MeSH heading. Second, topics provide a basis upon which we can compare MeSH headings, and compute quantities related to the MeSH hierarchy. A simple task is to explain differences in closely related MeSH headings. This is useful for educating PubMed users as to the distinctions between MeSH headings, and also suggesting other MeSH headings to use in searches. Below we list topics related to three MeSH headings related to cocaine, and two MeSH headings related to smoking:

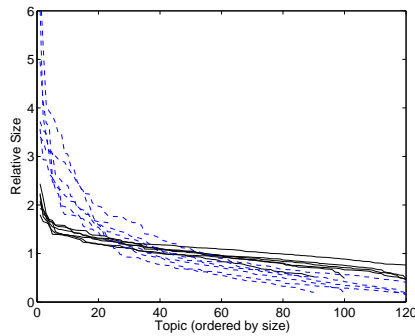


Fig. 1. Spectrum of topic size relative to $\frac{1}{T}$ for topic model (solid) and author-topic model (dashed), showing that the topic model produces a flatter distribution of topics for the six query results sets.

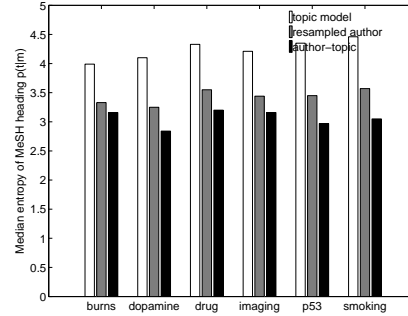


Fig. 2. Median entropy of MeSH heading topic distributions, showing that the author-topic model learns clearer associations with learned topics than the topic model, with the resampled author model results falling in between.

```

Cocaine-Related-Disorders [q=drug]
(0.32) [t114] cocaine user crack drug dependence abuse urine reported day cocaine-dependent ...
(0.05) [t66] drug drug-use substance-use substance substance-abuse drug-abuse illicit-drug alcohol ...
(0.04) [t6] treatment outcome program client outpatient residential abuse-treatment follow-up ...
Cocaine[q=drug]
(0.39) [t114] cocaine user crack drug dependence abuse urine reported day cocaine-dependent ...
(0.04) [t9] urine concentration positive hair sample testing morphine specimen detection test ...
(0.04) [t77] group subject n= found male individual finding examined evaluated test clinical ...
Crack-Cocaine [q=drug]
(0.38) [t114] cocaine user crack drug dependence abuse urine reported day cocaine-dependent ...
(0.07) [t12] sample likely less characteristic multiple recent demographic risk report similar ...
(0.05) [t97] sexual sex partner condom sexually std female transmitted women intercourse risk ...
- - -
Smoking [q=smoking]
(0.23) [t54] smoker smoking cigarette cigarette-smoking effect nonsmoker non-smoker smoked ...
(0.16) [t129] smoking smoker tobacco smoke consumption tobacco-use daily smoked cigarette ...
(0.05) [t108] year age change period young aged pattern related relationship rate ...
Tobacco-Use-Disorder [q=smoking]
(0.17) [t139] dependence measure scale negative addiction questionnaire score positive ...
(0.12) [t129] smoking smoker tobacco smoke consumption tobacco-use daily smoked cigarette ...
(0.09) [t147] nicotine cigarette effect gum smoker patch nrt mg level replacement ...

```

In all three cocaine-related headings, the most likely topic is [t114], capturing clear content related to cocaine use. The topics that follow give a clue to the distinction between these headings: Cocaine-Related-Disorders features [t66] (substance abuse) and [t6] (treatment), Cocaine features [t9] (testing), and Crack-Cocaine is further distinguished by its inclusion of [t97] (sex).

In the next example, Smoking includes a generic and shared tobacco smoking topic [t129], whereas the Tobacco-Use-Disorder is distinguished by topics [t139] (dependence/addiction) and [t147] nicotine. This usage of these two MeSH headings is consistent with the Annotation and Scope Notes provided in the Descriptor Data as shown in the MeSH browser⁴ which states: Smoking = In-

⁴ <http://www.nlm.nih.gov/MeSH/MBrowser.html>

haling and exhaling the smoke of tobacco; and Tobacco-Use-Disorder = Tobacco used to the detriment of a person's health.

3.1 Which MeSH headings are similar?

Knowing the topic distributions for MeSH headings allows us to compute the distance between two headings using the symmetric KL divergence, $KL^*(p(t|m_1)||p(t|m_2))$. This distance computation provides additional insight into the relationship between related MeSH headings. For example, the MeSH browser page for Substance-Related-Disorders mentions Street-Drugs (under See Also), but does not mention Urban Population or Psychotropic Drugs, which we computed as also being closely related to Substance-Related-Disorders. We display these connections in Figure 3, which shows connections that exist in the MeSH hierarchy (solid and dashed lines), as well as connections that are learned via topics (dotted lines). This type of visualization can immediately convey to a user the *actual* relationships between MeSH headings – possibly even surprising connections – as inferred from their pattern of usage in MEDLINE articles.

3.2 MeSH headings versus qualifiers

Qualifiers have an important role in the application of MeSH headings. Since there are only approximately 80 qualifiers and 25,000 MeSH headings, qualifiers provide a broad subject categorization that is somewhat orthogonal to that provided by the MeSH headings. Qualifiers can not appear on their own (they always qualify a MeSH heading), so can we learn topic distributions for qualifiers that span different query results sets? By treating qualifiers as additional separate MeSH headings, we learned the following distributions over topics for epidemiology, for the smoking and drug query results sets:

```
epidemiology [q=smoking]  
(0.15) [t103] prevalence population survey health high adult higher aged sample current ...  
(0.14) [t30] risk-factor age year incidence risk factor smoking population prevalence ...  
(0.07) [t95] year trend data change period health age increase national pattern ...  
(0.07) [t27] rate objective background higher likely analysis assess percentage total ...
```

```
epidemiology [q=drug]  
(0.17) [t61] prevalence survey national data year sample estimate age population aged ...  
(0.09) [t103] women men female male gender pattern higher lifetime sample respondent ...  
(0.09) [t21] likely year population less reported higher rate month status objective ...  
(0.08) [t64] risk age year history factor association male frequency greater female ...
```

These two topic distributions show a relatively consistent list of topics for the two query results sets. The top topic in each case is about prevalence, surveys and populations. Both sets also include a topic related to risk factors, and the remaining topics share several words that relate to rates, change over time, and likelihoods. Also note the virtual absence of words directly related to the original query. With the exception of the word 'smoking' appearing in [t30], there are no words that are about the query areas. This shows that the models are particularly good at learning a separate meaning for both MeSH headings, and their qualifiers.

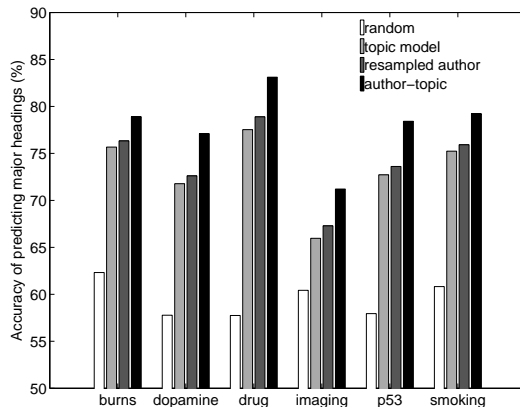


Fig. 4. Accuracy of predicting which MeSH headings are major. Random performs better than 50% because the number of major MeSH headings per document is used.

3.3 Predicting major MeSH headings

We have described several ways in which our topic representation is useful for explaining, interpreting and understanding MeSH headings. But how well do they perform on predictive tasks? We setup the following task: Given all the MeSH tags (major and minor) applied to a test article, predict which tags are major. For each unseen test article, we list by name all the MeSH tags, and indicate the number of major MeSH tags. For example, the article entitled “Effects of caffeine in overnight-withdrawn consumers and non-consumers” has major MeSH tags {Affect, Caffeine, Cognition} and minor MeSH tags {Adolescent, Adult, Attention, Female, Humans, Male, Placebos, Reaction Time, Saliva, Substance Withdrawal Syndrome}. Beyond some of the check-tags like ‘Human’, it is not immediately obvious (from just looking at the title) which tags would be major. We used the three models to rank MeSH headings in order of $p(m|d) = \sum_t p(m|t)p(t|d)$. The results, shown in Figure 4 show that all models have clear predictive ability that is better than random, with the author-topic model having the best accuracy.

4 Discussion and Conclusions

In this paper we have shown examples of how topic modeling is useful for interpreting and understanding MEDLINE’s MeSH subject headings. We start by using the standard topic model and the author-topic model, then introduce the resampled author model which is a hybrid of the two. Using these topic models we show how the learned distribution over topics for each MeSH heading is a useful representation for comparing and contrasting MeSH headings. We acknowledge that the learned topic interpretation of MeSH headings depends on

the query results set, however for sufficiently large query results sets we expect to learn relatively consistent interpretations of MeSH headings.

Previous studies that analyzed PubMed/MEDLINE usage using PubMed query logs analyzed statistics of PubMed users, their actions and their queries ([6, 7]). An analysis of query expansion using MeSH was reported in [1]. Topic models were applied to MEDLINE articles for the purpose of predicting MeSH headings in [8], and a similar semantic analysis of the WordNet hierarchy was conducted by [9]. The concept-topic models of [10] also relate learned topics to existing concept hierarchies, however that work focuses on tagging unseen documents using a mix of learned topics and existing concepts, in contrast to our focus of interpreting and understanding the existing hierarchy.

The topic modeling approach presented here has some useful and flexible features. While not explored in this paper, the topic models' word-level annotation of topics and authors (i.e. MeSH headings) could be valuable for annotating which sections of longer documents are most relevant to each MeSH tag applied. More broadly, this framework is generally useful for any collection that is tagged with subject headings. Extensions to this work could include devising topic models to validate subject heading hierarchies and creating a tool to support ontology maintenance.

Acknowledgement: NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program

References

1. Lu, Z., Kim, W., Wilbur, W.J.: Evaluation of query expansion using MeSH in PubMed. *Inf. Retr.* **12**(1) (2009) 69–80
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *Journal of Machine Learning Research* **3** (2003) 993–1022
3. Griffiths, T., Steyvers, M.: Finding scientific topics. In: *Proceedings of the National Academy of Sciences*. Volume 101. (2004) 5228–5235
4. Buntine, W., Jakulin, A.: Applying discrete PCA in data analysis. In: *UAI*. (2004) 59–66
5. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: *UAI*. Volume 20. (2004) 487–494
6. Herskovic, J., Tanaka, L., Hersh, W., Bernstam, E.: A day in the life of PubMed: Analysis of a typical day's query log. *J Am Med Inform Assoc* **14** (2007) 212–220
7. Lin, J., Wilbur, W.J.: Modeling actions of PubMed users with n-gram language models. *Information Retrieval* **12**(1) (2008) 69–80
8. Mörchen, F., Dejori, M., Fradkin, D., Etienne, J., Wachmann, B., Bundschuh, M.: Anticipating annotations and emerging trends in biomedical literature. In: *SIGKDD*, Las Vegas, Nevada (2008) 954–962
9. Snow, R., Jurafsky, D., Ng, A.Y.: Semantic taxonomy induction from heterogeneous evidence. In: *ACL*, Sydney, Australia (2006) 801–808
10. Chemudugunta, C., Smyth, P., Steyvers, M.: Combining concept hierarchies and statistical topic models. In: *CIKM*. (2008) 1469–1470