Understanding PubMed Search Results using Topic Models and Interactive Information Visualization

Zhiguo Yu

University of Texas Health Science Center at Houston, Zhiguo.Yu@uth.tmc.edu

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthshis_dissertations

Part of the Other Medicine and Health Sciences Commons

Recommended Citation
https://digitalcommons.library.tmc.edu/uthshis_dissertations/42
Understanding PubMed Search Results using Topic Models and Interactive Information Visualization

By

Zhiguo Yu, MS

APPROVED:

Todd R Johnson, PhD, Chair

Elmer Bernstam, MD, MSE

Trevor Cohen, MBChB, MD, PhD

Cui Tao, PhD

Byron C Wallace, PhD

Date approved: ___________________
Understanding PubMed Search Results Using Topic Models and Interactive Informative Visualization

A Dissertation

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
School of Biomedical Informatics
in Partial Fulfilment of the Requirements for the Degree of
Doctor of Philosophy

By

Zhiguo Yu, M.S.

University of Texas Health Science Center at Houston

2017

Dissertation Committee:

Todd Johnson, PhD¹, Advisor
Elmer Bernstam, MD, MSE¹,²
Trevor Cohen, MBChB, MD, PhD¹
Cui Tao, PhD¹
Byron Wallace, PhD³

¹The School of Biomedical Informatics, University of Texas Health Science Center at Houston
²McGovern Medical School, University of Texas Health Science Center at Houston
³College of Computer and Information Science, Northeastern University
Dedication

To those who have helped expand my world. I’m grateful forever for each one of you.
Acknowledgements

This work was supported in part by the UTHHealth Innovation for Cancer Prevention Research Training Program Predoctoral Fellowship (Cancer Prevention and Research Institute of Texas (CPRIT)). The content is solely the responsibility of the author and does not necessarily represent the official views of the CPRIT.

At the beginning, I would like to thank my academic advisor, Dr. Todd Johnson. This dissertation would be impossible without his knowledge, support, encouragement, and patience. He is not only a good advisor, teacher, role model, but also a friend that cares about your life and emotions. It’s been 6 years since I started working for him. His wisdom and spirit will always be my guidance in the future. Thank you. I’m also very grateful for my PhD committee, Dr. Elmer Bernstam, Dr. Trevor Cohen, Dr. Byron Wallace, and Dr. Cui Tao, who have been generous with their time, support, and insightful guidance to improve my work. In 2016, I had the great pleasure of spending a whole summer at National Library of Medicine at Lister Hill working with Dr. Olivier Bodenreider. I was truly impressed by his leadership and the way he conducts research, which is going to be a great example to follow in my future career.

I’d also take this opportunity to thank the faculty from the CPRIT innovation fellowship Dr. Roberta Ness, Dr. Patricia Mullen, Dr. David Loose, and Dr. George Delclos. These
mentors have taught me to be more innovative and accurate in research and presentations. Also I’m very grateful for their efforts, resources, and time on my English accent training. Besides, I’d also like to thank my CPRIT fellowship mates for their innovative suggestions, comments and feedback on my work.

Last but not least, I would like to express my deepest gratitude to my parents, my brother, my grandmother, and the rest of my family, for their endless love and support I have always received since I was born.
Abstract

With data increasing exponentially, extracting and understanding information, themes and relationships from larger collections of documents is becoming more and more important to researchers in many areas. PubMed, which comprises more than 25 million citations, uses Medical Subject Headings (MeSH) to index articles to better facilitate their management, searching and indexing. However, researchers are still challenged to find and then get a meaningful overview of a set of documents in a specific area of interest. This is due in part to several limitations of MeSH terms, including: the need to monitor and expand the vocabulary; the lack of concept coverage for newly developing areas; human inconsistency in assigning codes; and the time required to manually index an exponentially growing corpus. Another reason for this challenge is that neither PubMed itself nor its related Web tools can help users see high level themes and hidden semantic structures in the biomedical literature.

Topic models are a class of statistical machine learning algorithms that when given a set of natural language documents, extract the semantic themes (topics) from the set of documents, describe the topics for each document, and the semantic similarity of topics and documents. Researchers have shown that these latent themes can help humans better understand and search documents. Unlike MeSH terms, which are created based on important concepts throughout the literature, topics extracted from a subset of documents
are specific to those documents. Thus they can find document-specific themes that may not exist in MeSH terms. Such themes may give a subject area-specific set of themes for browsing search results, and provide a broader overview of the search results.

This first part of this dissertation presents the TopicalMeSH representation, which exploits the ‘correspondence’ between topics generated using latent Dirichlet allocation (LDA) and MeSH terms to create new document representations that combine MeSH terms and latent topic vectors. In an evaluation with 15 systematic drug review corpora, TopicalMeSH performed better than MeSH in both document retrieval and classification tasks. The second part of this work introduces the “Hybrid Topic”, an alternative LDA approach that uses a ‘bag-of-MeSH&words’ approach, instead of just ‘bag-of-words’, to test whether the addition of labels (e.g. MeSH descriptors) can improve the quality and facilitate the interpretation of LDA-generated topics. An evaluation of this approach on the quality and interpretability of topics in both a general corpus and a specialized corpus demonstrated that the coherence of ‘hybrid topics’ is higher than that of regular bag-of-words topics in both specialized and general corpora. The last part of this dissertation presents a visualization tool based on the ‘hybrid topics’ model that could allow users to interactively use topic models and MeSH terms to efficiently and effectively retrieve relevant information from tons of PubMed search results. A preliminary user study has been conducted with 6 participants. All of them agree that this tool can quickly help them understand PubMed search results and identify target articles.
Vita

2007-2011 ........................................ B.S., Electronic Engineering and Information Science, Beijing Institute of Technology, Beijing, China

2011-2013 ........................................ M.S., Electronical and Computer Engineering, University of Kentucky, Lexington, KY, USA

2014 to present ....................................... Graduate Research Assistant, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX

2014 to present ....................................... Pre-doctoral Fellow, PhD Student, School of Public Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX
Field of Study

Health Informatics
# Table of Contents

Dedication .................................................................................................................. ii

Acknowledgements ................................................................................................... iii

Abstract ..................................................................................................................... v

Vita .............................................................................................................................. vii

Table of Contents ..................................................................................................... ix

List of Tables ............................................................................................................. x

List of Figures .......................................................................................................... xi

Chapter 1: Introduction ............................................................................................. 1

Chapter 2: Related Work ........................................................................................ 11

Chapter 3: TopicalMeSH ....................................................................................... 14

Chapter 4: Hybrid Topics ....................................................................................... 37

Chapter 5: VizTM: Semantic Facet Browser using Topic Models ......................... 53

Chapter 6: Conclusion, Limitation, and Future Work ........................................... 71

References ............................................................................................................... 76

Appendix A: User Study Questions ...................................................................... 83

Appendix B: Information Consnet ....................................................................... 86

Appendix C: Letter of Information ...................................................................... 89
List of Tables

Table 1. Systematic drug review corpora description........................................... 23
Table 2. Relevant MeSH Terms for Drugs .......................................................... 24
Table 3. F-scores of support vector machines ...................................................... 30
Table 4. F-scores of logistic regression .............................................................. 31
Table 5. F-scores of decision tree ...................................................................... 32
Table 6. Comparison to tf-idf MeSH and sLDA.................................................... 33
Table 7. Topics generated based on ‘bag-of-MeSH&word’.................................. 42
Table 8. # of Topics with 0,1,n MeSH descriptors (n>1).................................... 51
Table 9. Details of participants’ queries ............................................................. 67
List of Figures

Figure 1. Correspondence matrix of MeSH terms and topics.......................... 20
Figure 2. A schematic to inducing TopicalMeSH representations of articles ....... 21
Figure 3. Individual precision and recall of 15 corpora................................. 28
Figure 4. Overall average precision and recall of 15 corpora.......................... 29
Figure 5. Comparison of mean TC-W2V topic coherence scores for different numbers of topics k, generated from the general corpus .................................................. 48
Figure 6. Comparison of mean TC-W2V topic coherence scores for different numbers of topics k, generated from the specialized corpus .................................................. 48
Figure 7. Plot of mean TC-W2V topic coherence scores for different numbers of topics k, generated from the general corpus .................................................. 50
Figure 8. Plot of mean TC-W2V topic coherence scores for different numbers of topics k, generated from the specialized corpus .................................................. 50
Figure 9. Main Page of VizTM .......................................................... 57
Figure 10. VizTM- Token Filter Function ............................................. 58
Figure 11. VizTM- Topic Filter Function .............................................. 59
Figure 12. VizTM- Topic Page .......................................................... 61
Figure 13. VizTM- Document Page ...................................................... 62
Chapter 1: Introduction

Knowledge discovery is a fundamental and important activity in biomedical research. Given that unstructured data is increasing exponentially, extracting and understanding information, themes, and relationships from large collections of documents becomes increasingly important for biomedical researchers (Holzinger and Jurisica, 2014). It is common for biomedical researchers to read and analyse published articles related with their research areas to get specific information and stay up-to-date. However, it is increasingly difficult for researchers to keep up with even narrowly defined research areas by reading publications (Fraser and Dunstan, 2010). PubMed is one of the most common tools for biomedical researchers to electronically search and retrieve biomedical literature. Although PubMed provides a broad, up-to-date and efficient search interface, it is still challenging for users to quickly identify information relevant to their specific information needs (Lu, 2011). This chapter introduces the information overload issue in PubMed, the utility of MeSH terms in PubMed, the method of topic models that I use to address the information overload problem, the relationships between MeSH and topic models, outlines the dissertation structure, and summarizes the contributions of this work.

1.1 - Information Overload in PubMed

As one of the largest repositories of biomedical literature, MEDLINE now indexes over 27 million citations, and is growing at a 4% annual rate (Lu, 2011). To better manage and search articles in MEDLINE, the US National Library of Medicine (NLM) developed the
Medical Subject Heading (MeSH) controlled vocabulary for indexing articles in MEDLINE. There are 27,883 descriptors in the 2016 MeSH, with over 87,000 entry terms that assist in finding the most appropriate MeSH descriptor (for example, ‘Vitamin C’ is an entry term to the MeSH descriptor ‘Ascorbic Acid’). In the 2016 MeSH, 82 qualifiers can be attached to MeSH descriptors to describe a particular aspect of a concept, such as ‘adverse effects’, ‘diagnosis’, etc. NLM has spent considerable effort and resources to create and maintain the MeSH vocabulary. Each year, the MeSH specialists revise and update the MeSH vocabulary to cover emerging research areas and improve indexing consistency and efficiency. MeSH specialists are responsible for areas of the health sciences in which they have knowledge and expertise. MEDLINE indexers make suggestions for new descriptors to MeSH specialists during their indexing processes. In addition, MeSH specialists also collect new terms as they appear in the scientific literature or emerging areas of research. After defining these terms within the context of the existing vocabulary, MeSH specialists recommend their addition to MeSH. During MEDLINE year-end processing (YEP) activities, changes made to MeSH are applied to MEDLINE (retrospectively) for conformance with the current version of MeSH. In 2000, the NLM launched its own indexing initiative project and developed the Medical Text Indexer (MTI) to assist human annotators with indexing recommendations in the form of MeSH headings and heading/subheading pairs (Aronson, Mork, Gay, and Humphrey, 2004). A relatively small group of indexing contractors and staff at the US NLM are responsible for attaching MeSH terms to biomedical articles. This semi-
automated indexing process may avoid many issues involved with natural language processing, but also has several limitations:

**MeSH Limitation 1:** Human consistency on MeSH indexing ranges from 33.8% to 74.7% depending on the type and category of indexing terms (Funk and Reid, 1983). Note that this finding predates the use of the MTI to assist MeSH specialists.

**MeSH Limitation 2:** MeSH terms in each abstract are binary tags. Beyond its major/regular MeSH assignment, there are no precise MeSH term weights for each article to indicate the relative strength of each assigned MeSH term.

**MeSH Limitation 3:** Missing MeSH terms: NLM's MeSH specialists continually revise and update the MeSH vocabulary. They define new terms as they appear in the literature or in emerging areas of research within the context of the existing vocabulary and recommend their addition to MeSH. There is thus a high chance of missing emerging concepts such as "Biomedical Informatics" or "big data".

**MeSH Limitation 4:** Some citations may be missing relevant MeSH terms. Given that MEDLINE is rapidly increasing in size, it is difficult to assign MeSH terms to citations immediately after they become searchable online. According to NLM's recent statistical analysis, 25% of the citations are completed within 30 days of receipt, 50% within 60 days, and 75% within 90 days (Huang, Névéol, and Lu, 2011). Furthermore, NLM does not typically re-index old articles with new MeSH terms (NLM, 2015).

**MeSH Limitation 5:** The utility of developing combinations of MeSH descriptors as an input query has been shown to be a valuable mechanism to identify
relevant articles, and subsequently possible new knowledge (Srinivasan, 2001; Srinivasan, 2004), but MeSH terms by themselves do not indicate meaningful clusters or themes in a specific subject area.

Despite the limitations of MeSH terms, PubMed uses MeSH terms as its one of the most important resources to help users retrieve relevant literature. One attractive feature of PubMed is its automatic query expansion process, called Automatic Term Mapping (ATM). ATM compares and maps untagged terms from the user query to lists of re-indexed terms in PubMed’s translation tables (in the following order): the MeSH table (mapping search terms to MeSH concepts), the journals translation table (mapping search terms to journal names), and the author index (mapping search terms to author names). Query expansion using MeSH in PubMed can generally improve retrieval performance, but the improvement may not affect end PubMed users in realistic situations (Lu, Kim, and Wilbur, 2009). Another problem is that PubMed displays search results as linear lists of items over many pages. However, most users only click the top ranked returned documents (e.g. the first 20 returned documents) to access more detailed information (Dogan, Murray, Névéol, and Lu, 2009).

With maturity in the field of text mining, an increasing number of web tools have been developed to provide comparable literature search service to PubMed with the aim of helping users quickly and efficiently search and retrieve relevant publications. While
critical to helping researchers, most tools (25 of 28 in Lu’s review) still present results as linear lists of items over many pages (Lu, 2011).

Even the best (most relevant first) linear list of results will not give the user a useful overview of the themes and semantic connections in the search results. For that, researchers have turned to visual overviews, such as LigerCat and MeSHy. LigerCat uses a MeSH term cloud to help users get an overview of search results and each MeSH term is used to filter and retrieve relevant results. MeSHy uses the frequencies of occurrence, concurrences, and the semantic similarities of MeSH terms in retrieved PubMed documents to create MeSH term pairs to help users better identify the most relevant information. However, both these tools still suffer from the limitations of MeSH terms described above. Hence, innovative methods combined with interactive information visualization is needed to help users get overall themes and understand the hidden semantic structure of large collections of documents. This dissertation addresses both of these needs by using MeSH terms and topic models to uncover semantic themes and relationships, along with a novel interactive information visualization of these themes and the corresponding documents.

1.2 - Topic Models

Topic modeling (Blei, 2012) is a popular technique for automatically summarizing and revealing the hidden structure of large collections of documents. Given a corpus of
natural language documents, topic models can extract the semantic themes (topics) within these and then describe the thematic composition of each document in the corpus, and hence the semantic similarity between topics and documents.

Although researchers continue to develop better and faster topic model algorithms (Teh, Jordan, Beal, and Blei, 2006; Blei and Lafferty, 2006; Dai, Wei, Zheng, Kim, Lee, Yin, Ho, and Xing, 2013), there is limited work on assessing their utility in real world applications. This is because topic model developers typically focus on intrinsic quality: that is, how well (probabilistically) the topic model explains (held-out) text. Such intrinsic measures do not necessarily reflect utility in practice (Chang, Gerrish, Wang, Boyd-graber, and Blei, 2009). Another reason is that making sense of topic models’ results for general users with proper visualizations is challenging (Blei, 2012). Topic Model Browser is a visualization tool that allows users to use topic model results to explore large collections of documents (Chaney and Blei, 2012). In our previous evaluation (Yu, Johnson, and Kavuluru, 2013), we invited three biomedical researchers from different research domains to evaluate the browser. Though researchers found this tool interesting and helpful to use, they pointed out that the browser could be improved in several ways, such as filtering with a combination of topics, mapping from words/phrases to related topics, etc.

In this dissertation, I introduce an innovative visualization tool that could allow users to interactively use topic model to more efficiently and effectively retrieve, explore, and
understand PubMed search results. This tool is motivated by specific information needs from our previous evaluation (Yu, Johnson, and Kavuluru, 2013). However, PubMed articles are already indexed using MeSH terms. One crucial question before applying topic models to PubMed is whether topic models can bring additional useful utility over MeSH.

1.3 – MeSH indexing Versus Topics

MeSH terms are manually created and maintained by domain experts to cover all generally important themes. However, topic models are used to extract semantic themes that are specific to a subset of documents. Topic models can find corpus-specific themes that may not exist in MeSH. Such themes may uncover a specific set of topics for a particular domain or sub-domain, thus potentially providing a better overview of search results (Yu, Johnson, and Kavuluru, 2013). Given the difference between topics and MeSH terms, we formulate the following two main hypotheses:

**Hypothesis 1:** Topic models can improve the utility of MeSH terms.

**Hypothesis 2:** Adding MeSH terms to topic models’ “bag-of-words” assumption can improve the quality and facilitate the interpretation of generated topics.

Topic models have shown great ability to atomatically handle large collections of documents. Given the MeSH limitations listed in Chapter 1.1, we would like to answer the following questions in our designed experiments:
**Question 1:** Can we use topic models to identify new themes that may not be in the MeSH vocabulary? (MeSH Limitation 3)

**Question 2:** Can we use topic models to help identify citations that are missing relevant MeSH terms (MeSH Limitation 1 and 4)?

**Question 3:** Can we use topic models to assign each MeSH term a precise weight in each citation (MeSH Limitation 2)?

**Question 4:** Can we use topic models to identify clusters of MeSH terms that could better represent users’ specific information needs (MeSH Limitation 5)?

For the first hypothesis, we test how much topic models can bring to PubMed over MeSH. For this purpose, we introduce a novel representation (TopicalMeSH) that combines both topic model’s unsupervised learning results and MeSH’s human curated indexing. With this novel representation, questions 2 and 3 are addressed in this experiment. In the second hypothesis, we test how much MeSH terms can help improve the quality and facilitate the interpretation of topic model results. For this purpose, we introduce an alternative topic model approach (Hybrid Topics) by changing the standard “bag-of-words” assumption to “bag of MeSH and Words”. To address questions 1 and 4, we analyze hybrid topics in two different contexts: General Corpus and Specialized Corpus. To help human better understand and explore tons of PubMed search results, we developed a visualization tool based on the “Hybrid Topic” model that could allow users to interactively use both topic models and MeSH terms to efficiently and effectively retrieve relevant information. The visualization is designed following Shneiderman’s visualization mantra “Overview first, filter and zoom, details on demand ”.
One of the challenges with making practical use of topic models is that users often do not understand what each topic “means” since the topics are distributions over words, not a single understandable word or phrase that summarizes the concept. TopicalMesh effectively solves this problem by associating topics with corresponding MeSH terms. This provides a convenient and familiar terminology to present to users.

1.4- Summary of Contributions

The following are the main contributions presented in this dissertation:

**Contribution 1:** Developed and evaluated a novel representation (TopicalMeSH) for documents that utilizes both discrete human curated document tags (e.g., MeSH descriptors) and vector representations learned through unsupervised machine learning algorithms (e.g., topic models). TopicalMeSH can assess each MeSH term’s weight for each document and also (implicitly) assign MeSH terms to documents that may be missing some relevant MeSH terms. In this way, TopicalMeSH not only improves such downstream applications using MeSH, but also provides a convenient and familiar terminology to help users interpret topics.

**Contribution 2:** Introduced an alternative LDA approach by including discrete human curated tags to its original “bag-of-words” assumption to improve the quality and facilitate the interpretation of topic model results. This addition of human curated tags improve the coherence of topics and helped humans better interpret topic model results.
Contribution 3: Developed an innovative interactive visualization tool that could allow users to more efficiently and effectively understand and explore PubMed search results. The user study of this tool helped us better understand the human-computer interaction issues around using topic models. It also demonstrated how these models can transform biomedical data in scientific abstracts into information.

1.5- Dissertation Structure

The remainder of this dissertation is laid out in the following order. In Chapter 2, I review prior work related to MeSH terms, topic models, and PubMed web tools. In Chapter 3, I present the novel TopicalMeSH representation for PubMed citations and show how TopicalMeSH improves the utility of MeSH terms. In Chapter 4, I describe the hybrid topics that utilize MeSH descriptors to help improve the quality and facilitate the interpretation of topic models. In Chapter 5, I introduce the visualization tool that uses the hybrid topic model to allow users to interactively use topic models and MeSH descriptors to explore and understand PubMed search results. In Chapter 5, I also present an evaluation of VizTM with real world PubMed queries to demonstrate its ability to help users efficiently and effectively explore and understand large collections of documents. The dissertation then concludes by describing its significance, contributions to the field of biomedical informatics and biomedical research, and directions for future research.
Chapter 2: Related Work

This chapter gives a summary of the current research related to topic models, PubMed and MeSH.

2.1- Topic Models in Biomedical Informatics

Topic models have broad applicability in biomedical informatics. Recent examples include treatment discovery from clinical cases (Zhang, et al., 2011; Yao, et al., 2014), predicting behaviour codes from couple therapy transcripts (Atkins, et al., 2012), risk stratification in ICU patients from nursing notes (Lehman, et al., 2012), summarizing themes in large collections of clinical reports (Arnold, et al., 2015), and discovering health related topics in social media (Lin, et al., 2010; Paul and Dredze, 2014; Wang, Agichtein, and Benzi, 2012).

2.2- Topic Model with MeSH Terms

Several studies have examined the application of topic models to PubMed data. Most recently, Zhu et al. (Zhu et al., 2009) used labeled Latent Dirchlet Allocation (labeled LDA) to annotate MeSH terms. Labeled LDA is a supervised topic model for uncovering latent topics that correlate with user tags in labeled corpora. However, labeled LDA did not perform better than the other methods evaluated (Zhu, et al., 2013). Elsewhere, Newman et al. (Newman, Karimi, and Cavedon, 2009) presented a resampled author
model that combines both general LDA and the author-topic model (in this case MeSH terms were used as the “authors”). The resampled author model provided an alternative and complementary view of the relationships between MeSH terms. However, their model did not outperform the author-topic model with respect to predicting MeSH terms for biomedical articles.

MeSH’s hierarchical structure has also been explored to identify biomedical topic evolution (He, 2012). More recently, the graph-sparse LDA model was developed to generate more interpretable topics by leveraging relationships expressed by controlled vocabulary structures. In this model, a few concept-words from the controlled vocabulary can be identified to represent generated topics. MeSH was shown to work well in this model to help summarize biomedical articles (Doshi-Velez, 2015).

To the best of our knowledge there have been no previous evaluations of the correspondence between topics from topic models and MeSH terms in PubMed. Nor has there been an investigation into whether topics might complement MeSH terms.

2.3- Visualizations for PubMed

PubMed provides users a broad, up-to-date and efficient search interface, but users find it more and more challenging to quickly identify relevant information to their information needs given the exponentially growing biomedical literatures. Users are often overloaded by the list of results spread over many pages. In 2011, Lu evaluated 28 PubMed web
tools (Lu, 2011). 5 of them cluster results into topics given words, MeSH terms and MeSH structures. 25 of them still represent results as a linear list over many pages. MeSH terms are often used to help understand PubMed search results. LigerCat (Sarkar, Schenk, Miller, and Norton, 2009) uses a “MeSH cloud” to present an overview of search results. Each MeSH term in the cloud can be used as a filter to narrow down relevant results. MeSHy (Theodosiou, Vizirianakis, Angelis, Tsaftaris, and Darzentas, 2011) was developed to extract and rank MeSH pairs from PubMed search results to help users better discover unanticipated information.

There are also a few visualizations developed for PubMed search results using topic models. In 2015, Kirti Kamboj used nodes-and-edges representations to display LDA’s results and PageRank to uncover the importance and influence among topic networks. However, MeSH and content information of the citations were not included. Other visualizations of topic models (Chaney and Blei, 2012; Ganesan, Brantley, Pan, and Chen, 2015; Murdock and Allen, 2015; Hinneburg, Preiss, and Schröder, 2012; Gardner, et al., 2010; Smith, et al., 2014; Guille, et al., 2016) displays hidden relationships among documents and topics and even provide topic quality check. These tools can all be applied to PubMed, but none of them provide topics and words combination filters for relevant documents derived from real users information needs from PubMed search results (Yu, Johnson, and Kavuluru, 2013).
Chapter 3: TopicalMeSH

This chapter introduces the TopicalMeSH representation for PubMed citations, a combination of topic model results and MeSH terms, to address MeSH term’s limitations and to further improve document retrieval and document classification performance.

Chuang et al. (Chuang et al., 2009) developed an evaluation framework to assess the quality of latent topics generated by a topic model with respect to a reference set of topics independently developed by domain experts. This framework produces a correspondence matrix of similarity scores between the reference topics and the latent topics. Chuang et al.’s reference topics were derived from manually curated expert topic titles, key phrases, and sets of related documents. These were used to create word distributions for each reference topic using \( tf-idf \) weighting and normalization. In this section, I propose a variation of Chuang et al.’s method to compute the correspondence between MeSH terms and topics generated by LDA. The MeSH vectors are constructed in a similar way as Chuang et al. constructed their reference topics. I did not need a separate set of experts to create and assign key phrases, because MeSH terms already play that role for PubMed articles. Rather than using this correspondence to evaluate the quality of estimated topics (as per the original idea), here I use correspondence to induce a new representation that captures correlations between MeSH terms and latent topics.

This section is presented in the following order: a brief description of LDA; the construction of the correspondence matrix between topics and MeSH terms; the design of
my workflow, a brief description of the test data and evaluation measures; results, section summary and discussion.

3.1-Latent Dirichlet Allocation (LDA)

LDA (Blei, Jordan, and Ng, 2003) is a probabilistic model that assumes that each document is generated from a mixture of topics, and that each topic corresponds to a distribution over all words in the corpus. Informally, the ‘generative story’ for LDA is as follows. First, a document is generated by drawing a mixture of topics that the document is about. To generate each word in this document, one draws a topic from this distribution and subsequently selects a word from the distribution over the vocabulary of the whole corpus corresponding to this topic. The LDA algorithm uses this generative model to uncover the latent topics contained within a given a corpus. Specifically, it estimates the parameters that define document topic mixtures and the conditional probabilities of each word given each topic. Parameter estimation is usually done via sampling approaches.

The number of topics produced by LDA must be prespecified. Determining the ‘right’ number of topics for different data sets remains a challenge. When the number of topics increases, redundant and nonsense topics may be generated. When the number of topics is small, we may miss some good topics.
3.2- Correspondence between Topics and MeSH

Each topic generated by LDA is a distribution over all of the unique words in the corpus. To compute the similarity scores between inferred topics and (observed) MeSH terms, we represented each MeSH term as a distribution of the words contained in the documents to which it had been assigned. To create the distribution for a MeSH term, we collected all of the words in documents from the corpus that were tagged with that MeSH term. After removing PubMed’s stop words, we used \( tf-idf \) (Eq. (1)) to re-weight the remaining words in the documents tagged with that MeSH term and then normalized the resulting MeSH vector representations to sum to one.

\[
 tf - idf_{w,d} = tf_{w,d} \cdot \log \frac{N}{df_{w,D}} \tag{1}
\]

In Eq. (1), \( tf_{w,d} \) is the term frequency of word \( w \) in document \( d \), \( df_{w,D} \) is the document frequency that word \( w \) appears in all documents \( D \), and \( N \) is the total number of documents.

The dimensionality of our correspondence matrix is \( T \) by \( M \), where \( T \) is the number of topics in the topic model and \( M \) is the number of unique MeSH terms. Each entry of this matrix is a similarity score between the word frequency vectors constructed for a given MeSH term and a topic’s estimated distribution over words.
There are several ways to compute the similarity between two distributions, including cosine similarity, Spearman’s rank correlation coefficient, and the rescaled dot product. Chuang et al. (Chuang, et al., 2009) compared several different methods for computing the similarity between two sets of topics generated from two LDA models on the same corpus and found that the rescaled dot product best predicted human judgment of similarity in terms of precision and recall.

This product is computed for two vectors $P$ and $Q$ as follows:

\[
\text{Rescaled Dot Product} \equiv \frac{P \cdot Q - d_{\text{Min}}}{d_{\text{Max}} - d_{\text{Min}}}; \quad d_{\text{Max}} = \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \right}
Fig. 1 provides an illustrative correspondence matrix. Here, Topic 2 is the corresponding topic for MeSH 2. Note that a single MeSH term may have multiple corresponding topics and a single topic may also be mapped to multiple MeSH terms.

<table>
<thead>
<tr>
<th></th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>MeSH 1</td>
<td>0.8</td>
<td>0.2</td>
<td>0.7</td>
<td>0.4</td>
<td>0.1</td>
<td>...</td>
</tr>
<tr>
<td>MeSH 2</td>
<td>0.4</td>
<td>0.9</td>
<td>0.2</td>
<td>0.1</td>
<td>0.2</td>
<td>...</td>
</tr>
<tr>
<td>MeSH 3</td>
<td>0.1</td>
<td>0.7</td>
<td>0.3</td>
<td>0.8</td>
<td>0.2</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Fig. 1** - Correspondence matrix of MeSH terms and topics.

### 3.3- Workflow

Fig. 2 shows the approach we took in this work. Process MeSH generated two matrices: $M1$ and $M2$. $M1$ is the binary Document-to-MeSH matrix, which has a dimensionality of $D$ by $M$, where $D$ is the total number of documents and $M$ is the number of unique MeSH terms. $M2$ is the $M$ by $W$ MeSH-to-Word matrix, where $W$ is the number of unique words. To make it consistent with LDA’s distribution over words, we removed both PubMed’s stop words and low frequency words that only appear in one document. The latter words are of little use in LDA because it is sensitive to word co-occurrence. We used $tf-idf$ to weight the remaining words and then normalized the weights so that they summed to one.
Run Topic Model used the LDA-c code (http://www.cs.princeton.edu/blei/lda-c/index.html) to perform topic modeling using variational inference (Blei, Ng, and Jordan, 2003). Since our goal was to compare the utilities of MeSH terms and topic models, we simply used the number of unique MeSH terms \( M \) indexed in Input Data as the number of topics for LDA. We selected this number based on experiments (described below) that varied the number of topics for the test corpora.

**Fig. 2** - A schematic to inducing TopicalMeSH representations of articles.

The symmetric Dirichlet prior parameter, \( \alpha \), was set to \( 1/T \), where \( T \) is the number of topics. We set the convergence criterion for variational expectation maximization (EM) to 0.0001 and the maximum number of iterations of variational EM to 1000. In our experiments, the convergence criterion was met across all corpora within 70 iterations.

Running LDA generated two Matrices: \( M3 \) and \( M4 \). \( M3 \) is the Topic-to-Word matrix with dimensions \( T \) by \( W \). Finally \( M4 \) is the \( D \) by \( T \) Document-to-Topic matrix.
Compute Correspondence Matrix uses the rescaled dot product defined above to calculate the similarity score of each MeSH term and topic pair based on the $M_2$ and $M_3$ matrices. $M_5$ is the Topic- to-MeSH correspondence matrix, which has dimensionality $T$ by $M$. Recall that we set $T = M$ when running LDA. Next, we calculated the matrix product of $M_5$ and $M_4$ to get the $D$ by $M$ matrix $M_6$, which is the same dimensionality as the Document-to-MeSH matrix, $M_1$. Because $M_6$ included both topic model and Topic-to-MeSH correspondence information, we named $M_6$ our Document-to-TopicalMeSH matrix. Here we propose using this TopicalMeSH document representation for text mining tasks. Specifically, for our Evaluation, we evaluate the utility of the TopicalMeSH feature matrix $M_6$ compared to the original MeSH features in Matrix $M_1$ on two standard tasks: document retrieval and text classification.

### 3.4- Test Data and Evaluation

For both tasks, we used 15 publicly available systematic drug review corpora (Cohen, et al., 2006). Each corpus is a set of PubMed titles and abstracts for systematic literature reviews comparing classes of drugs used for treating specific conditions. The datasets include queries for randomized controlled trials by combining terms for health conditions and interventions with research methodology filters for therapies. Systematic reviewers read the titles and abstracts to assess which articles likely met the inclusion criteria for the corresponding review. This process is called citation screening, and is a laborious step.
in the systematic review process (Wallace, et al., 2010). The articles deemed relevant by systematic reviewers during this citation screening process constitute the positive instances for each corpus. Here we use only the PubMed citation data, not the full text articles. More details about these data are available in (Cohen, et al., 2006).

Table 1 provides a brief description of the 15 drug review corpora that we used here. For the number of unique MeSH terms, we took only MeSH descriptors into consideration and count only those MeSH Terms indexed in more than 10 documents.

Table 1

*Systematic drug review corpora description.*

<table>
<thead>
<tr>
<th>Drug Review Name</th>
<th>No of Articles</th>
<th>% of judged relevant articles</th>
<th>No. of Unique MeSH Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE Inhibitors</td>
<td>2544</td>
<td>1.6</td>
<td>333</td>
</tr>
<tr>
<td>ADHD</td>
<td>851</td>
<td>2.4</td>
<td>155</td>
</tr>
<tr>
<td>Antihistamines</td>
<td>310</td>
<td>5.2</td>
<td>57</td>
</tr>
<tr>
<td>Atypical Antipsychotics</td>
<td>1120</td>
<td>13</td>
<td>155</td>
</tr>
<tr>
<td>Beta Blockers</td>
<td>2072</td>
<td>2</td>
<td>336</td>
</tr>
<tr>
<td>Calcium Channel Blockers</td>
<td>1218</td>
<td>8.2</td>
<td>197</td>
</tr>
<tr>
<td>Estrogens</td>
<td>368</td>
<td>21.7</td>
<td>79</td>
</tr>
<tr>
<td>NSAIDS</td>
<td>393</td>
<td>10.4</td>
<td>80</td>
</tr>
<tr>
<td>Opiods</td>
<td>1915</td>
<td>0.8</td>
<td>273</td>
</tr>
<tr>
<td>Oral Hypoglycemics</td>
<td>503</td>
<td>27</td>
<td>90</td>
</tr>
<tr>
<td>Proton Pump Inhibitors</td>
<td>1333</td>
<td>3.8</td>
<td>165</td>
</tr>
<tr>
<td>Skeletal Muscle Relaxants</td>
<td>1643</td>
<td>0.5</td>
<td>236</td>
</tr>
<tr>
<td>Statins</td>
<td>3465</td>
<td>2.5</td>
<td>447</td>
</tr>
<tr>
<td>Triptans</td>
<td>671</td>
<td>3.6</td>
<td>78</td>
</tr>
<tr>
<td>Urinary Incontinence</td>
<td>327</td>
<td>12.2</td>
<td>55</td>
</tr>
</tbody>
</table>

*Note.* ACE = angiotensin-converting enzyme; ADHD = attention-deficit/hyperactivity disorder; NSAIDs = nonsteroidal anti-inflammatory drug.
Since the LDA method used here required the number of topics to be prespecified, we conducted experiments with various numbers of topics ranging from 5 to 300. Performance in the information retrieval task was quite consistent. Runs with larger numbers of topics had slightly better performance than those with smaller numbers of topics. Hence here we used the number of unique MeSH terms (a relatively high number for LDA in general) in the test corpus as the number of topics. Although it may seem odd that a topic model of only 5 topics did not dramatically lower performance, recall that each of the 15 corpora were based on a specific query tuned to the medication class of interest. Thus one dominant theme in each corpus is likely to be the targeted medication class. As a result, it is likely that a topic model, even one with only 5 topics, will directly identify this theme, highlighting an advantage of TopicalMeSH.

For the document retrieval task, we compared TopicalMeSH and MeSH’s performance on retrieving relevant documents in each of the 15 drug review corpora. We mapped the drug names to relevant MeSH terms. Table 2 reports the details of drug names and their relevant MeSH Terms. We used the online MeSH Browser
(http://www.nlm.nih.gov/mesh/2015/mesh_browser/MBrowser.html) to find each drug’s relevant MeSH terms first. Each drug may have multiple relevant MeSH terms and some MeSH terms may have several children terms under the MeSH tree structure. For each drug, we then picked the single term that achieved the best retrieval performance (measured using F-score) from its relevant MeSH terms and their children terms if any. For example, ‘Antihistamines’ can be mapped to MeSH term ‘Histamine Antagonists’
using MeSH browser. ‘Histamine Antagonists’ has several children terms, ‘Histamine H1 Antagonists’, ‘Histamine H2 Antagonists’, and ‘Histamine H3 Antagonists’. ‘Histamine H1 Antagonists’ achieved a higher F-score than ‘Histamine Antagonists’ based on the ‘Antihistamines’ corpus. Hence we choose ‘Histamine H1 Antagonists’ as its relevant MeSH term for this document retrieval task.

Table 2

Relevant MeSH Terms for Drugs.

<table>
<thead>
<tr>
<th>Drug Review Name</th>
<th>Relevant MeSH</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE Inhibitors</td>
<td>Angiotensin-Converting Enzyme Inhibitors</td>
</tr>
<tr>
<td>ADHD</td>
<td>Attention Deficit Disorder with Hyperactivity</td>
</tr>
<tr>
<td>Antihistamines</td>
<td>Histamine H1 Antagonists</td>
</tr>
<tr>
<td>Atypical Antipsychotics</td>
<td>Antipsychotic Agents</td>
</tr>
<tr>
<td>Beta Blockers</td>
<td>Adrenergic beta-Antagonists</td>
</tr>
<tr>
<td>Calcium Channel Blockers</td>
<td>Calcium Channel Blockers</td>
</tr>
<tr>
<td>Estrogens</td>
<td>Estrogen Replacement Therapy</td>
</tr>
<tr>
<td>NSAIDS</td>
<td>Anti-Inflammatory Agents</td>
</tr>
<tr>
<td>Opioids</td>
<td>Analgesics</td>
</tr>
<tr>
<td>Oral Hypoglycemics</td>
<td>Hypoglycemic Agents</td>
</tr>
<tr>
<td>Proton Pump Inhibitors</td>
<td>Proton Pump Inhibitors</td>
</tr>
<tr>
<td>Skeletal Muscle Relaxants</td>
<td>Muscle Relaxants</td>
</tr>
<tr>
<td>Statins</td>
<td>Hydroxymethylglutaryl-CoA Reductase Inhibitors</td>
</tr>
<tr>
<td>Triptans</td>
<td>Sumatriptan</td>
</tr>
<tr>
<td>Urinary Incontinence</td>
<td>Urinary Incontinence</td>
</tr>
</tbody>
</table>

To reduce matrix sparsity, we used only MeSH main headings. We compared those MeSH terms’ document retrieval performance using the rows of matrix $M_1$ and those of $M_6$ in Fig. 2. $M_1$ records whether or not each MeSH term has been assigned to each
document, so those documents already annotated with the MeSH term (or terms) used in the query will be retrieved as relevant. M6 provides a mapping between terms in the MeSH vocabulary and the topics inferred by LDA, such that those documents in which topics corresponding to the MeSH term (or terms) in the query were ranked highest. The salient question was whether using the latter improves performance over the corresponding MeSH term. Although most real-world queries typically use more than one MeSH term or a combination of MeSH terms and other features, we used only the corresponding MeSH term, because our goal was to determine whether topic models could improve the utility of MeSH terms.

For the MeSH representation, we calculated a single pair of precision and recall scores (Eqs. (3) and (4)) based on the binary matrix $M1$. For the TopicalMeSH representation, we ranked those documents based on TopicalMeSH’s weights in matrix $M6$ and drew a precision-recall curve, in which the recall ranges from 10% to 100%. We computed both overall average and individual precision and recall for these 15 corpora. The overall average is just the arithmetic mean of the 15 individual precisions and recalls for these corpora.

In the classification task, we applied supervised machine learning methods to the corpus in an effort to learn to classify documents in the data sets with respect to their inclusion (1) or exclusion (-1) during citation screening. We used 80% of the documents in the data sets as training data and 20% as test data. We then evaluated three models that leveraged
different document representations as feature vectors: a linear kernel Support Vector Machine (SVM), logistic regression with an L2 penalty for coefficient regularization, and a decision tree. Using these models, we compared the following five different representations: MeSH, TopicalMeSH, Words, MeSH + Words, TopicalMeSH + Words. The Words representation used the ‘bag of words’ assumption with each word reweighted using the *tf-idf* measure (Eq. (1)). These three machine learning algorithms are implemented in the Python machine learning package *scikit-learn* (Pedregosa, et al., 2011). For the linear kernel SVM, we gave more importance to the relevant documents class with the class-weight setting as \(1:c\), where \(c\) is the proportion of the irrelevant documents to the relevant in each corpus. For the logistic regression, we used the L2 penalty with the regularization parameter \(C\) ranging from 0.01 to 1000 with the exponent base of 10. We report the best results. For the decision tree, we used the default settings.

In addition, we also compared TopicalMeSH’s performance in the classification task to the topic representation (Matrix 4 in Fig. 2) generated from LDA (Topic), *tf-idf* weighted MeSH, and a combination of topic representation and MeSH representation (Topic + MeSH) using SVMs. We also compared supervised LDA’s (sLDA) (Mccauliffe and Blei, 2008) document classification performance. sLDA is a supervised topic model for labeled documents, in which a response variable (label) associated with each document is also used in the generative model. For sLDA, we used 80% of the documents in the data sets to train sLDA and then used the trained model to infer the rest of the documents’ labels (relevant or irrelevant). We used the same number of topics in sLDA as we used in LDA.
We set the convergence criteria for variational EM to 0.0001 and the maximum number of iterations of variational EM to 70. And we also set the L2 penalty in sLDA to 0.01. In our experiments, the convergence criteria was met across all corpora within 70 iterations. For Topic and Topic+MeSH representations, we used SVMs as the classifier.

For all models, we used stratified 5-fold validation to assess performance. We used 5-fold (rather than 10-fold or leave-one-out cross fold validation) as a practical means to mitigate compute time: recall that we have 15 corpora and most of them have more than 1000 documents. We also use 3 classification methods. Within each method, we tested 5 different representations. For each of these, we trained and evaluated the machine learning algorithms with each feature set using the other four folds as training data. We averaged performance over these five folds to assess performance. For this task, we used the following evaluation metrics: *precision* (P), *recall* (R), and balanced *F-score* (F1). These measures are defined as:

\[
P = \frac{TP}{TP + FP} \quad (3)
\]

\[
R = \frac{TP}{TP + FN} \quad (4)
\]

\[
F1 = \frac{2 \cdot P \cdot R}{P + R} \quad (5)
\]

where TP, FP and FN denote true positive, false positive and false negative counts, respectively.
Fig. 3 - Individual precision and recall of 15 corpora.
3.5- Results

Result of document retrieval task

Fig. 3 provides graphical comparisons between the precision-recall curve of TopicalMeSH and the single precision-recall point of MeSH for these 15 corpora. TopicalMeSH achieved higher precision scores in 11 of 15 corpora with the same recall as achieved using MeSH. In addition, in 14 out of 15 corpora TopicalMeSH can achieve better recall while achieving precision that is similar to MeSH.
Fig. 4 provides a comparison of the overall average precision and recall between MeSH and TopicalMeSH over these 15 corpora. We can see that MeSH is clearly under TopicalMeSH’s precision-recall curve. The improvement of overall precision from MeSH to TopicalMeSH is about 5% (absolute) with the same recall of MeSH’s.

Result of classification task

To measure classification performance, we used F-score (Eq. (5)), which is the harmonic mean of precision and recall. We note that in practice, one would be much more concerned with recall than precision for the task of automated citation screening, as one aims to be comprehensive when conducting systematic reviews (Wallace, et al., 2010).

However, here we are using this as simply an illustrative biomedical text classification task; our focus is not specifically on the citation-screening problem. Tables 3 and 4 show the results of comparisons among the representations, MeSH, TopicalMeSH, Words, MeSH + Words, TopicalMeSH + Words, using the SVM, logistic regression, and the decision tree models, respectively. Table 6 shows the results of comparisons among the Topics, Topics + MeSH, and TopicalMeSH representations (all using SVM), as well as sLDA.

Table 3

F-scores of support vector machines.

<table>
<thead>
<tr>
<th>Drug Review Name</th>
<th>MeSH</th>
<th>TopicalMeSH</th>
<th>tf-idf Words</th>
<th>MeSH+Words</th>
<th>TopicalMeSH+Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE Inhibitors</td>
<td>9.28%</td>
<td><strong>33.20%</strong></td>
<td>20%</td>
<td>17.70%</td>
<td><strong>25.70%</strong></td>
</tr>
<tr>
<td>Drug Review Name</td>
<td>MeSH</td>
<td>TopicalMeSH</td>
<td>Word Embedding</td>
<td>MeSH+Words</td>
<td>TopicalMeSH +Words</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-------</td>
<td>-------------</td>
<td>----------------</td>
<td>------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>ACE Inhibitors</td>
<td>6.20%</td>
<td>18.80%</td>
<td>17.20%</td>
<td>13.70%</td>
<td>23.10%</td>
</tr>
<tr>
<td>ADHD</td>
<td>49.10%</td>
<td>44.60%</td>
<td>19.80%</td>
<td>56.90%</td>
<td>50.90%</td>
</tr>
<tr>
<td>Antihistamines</td>
<td>56.20%</td>
<td>27.10%</td>
<td>24.20%</td>
<td>53.90%</td>
<td>32.30%</td>
</tr>
<tr>
<td>Atypical Antipsychotics</td>
<td>24.70%</td>
<td>52.40%</td>
<td>54.50%</td>
<td>43.50%</td>
<td>57.20%</td>
</tr>
<tr>
<td>Beta Blockers</td>
<td>20.60%</td>
<td>27.10%</td>
<td>25.10%</td>
<td>27.10%</td>
<td>33.70%</td>
</tr>
<tr>
<td>Calcium Channel Blockers</td>
<td>16.60%</td>
<td>45.50%</td>
<td>40.20%</td>
<td>33.90%</td>
<td>48.50%</td>
</tr>
<tr>
<td>Estrogens</td>
<td>25.70%</td>
<td>45.10%</td>
<td>15.10%</td>
<td>30.70%</td>
<td>34.40%</td>
</tr>
<tr>
<td>NSAIDS</td>
<td>26%</td>
<td>55.70%</td>
<td>38%</td>
<td>34.10%</td>
<td>53.60%</td>
</tr>
<tr>
<td>Opioids</td>
<td>12.30%</td>
<td>6.70%</td>
<td>0</td>
<td>3.30%</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note.* Boldface numerals on the left show the best results for TopicalMeSH vs. MeSH, whereas those on the right highlight the best performance for MeSH + words vs. TopicalMeSH + words.
Table 5

*F*-scores of decision tree.

<table>
<thead>
<tr>
<th>Drug Review Name</th>
<th>MeSH</th>
<th>TopicalMeSH</th>
<th>tf-idf Words</th>
<th>MeSH+Words</th>
<th>TopicalMeSH +Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE Inhibitors</td>
<td>5%</td>
<td>17.50%</td>
<td>22.20%</td>
<td>22.90%</td>
<td>20.50%</td>
</tr>
<tr>
<td>ADHD</td>
<td>46.30%</td>
<td>38.50%</td>
<td>39.70%</td>
<td>49.50%</td>
<td>41.70%</td>
</tr>
<tr>
<td>Antihistamines</td>
<td>39.20%</td>
<td>38.50%</td>
<td>47.90%</td>
<td>48.50%</td>
<td>44.80%</td>
</tr>
<tr>
<td>Atypical Antipsychotics</td>
<td>34.50%</td>
<td>42.70%</td>
<td>47.20%</td>
<td>45.80%</td>
<td>46.20%</td>
</tr>
<tr>
<td>Beta Blockers</td>
<td>17.50%</td>
<td>19.70%</td>
<td>29%</td>
<td>30.20%</td>
<td>30.40%</td>
</tr>
<tr>
<td>Calcium Channel Blockers</td>
<td>26.80%</td>
<td>41.20%</td>
<td>43.50%</td>
<td>42.30%</td>
<td>43.10%</td>
</tr>
<tr>
<td>Estrogens</td>
<td>24.30%</td>
<td>34.40%</td>
<td>37.10%</td>
<td>39.60%</td>
<td>39%</td>
</tr>
<tr>
<td>NSAIDS</td>
<td>20.90%</td>
<td>42.20%</td>
<td>57.70%</td>
<td>56.20%</td>
<td>53.10%</td>
</tr>
<tr>
<td>Opioids</td>
<td>5%</td>
<td>9%</td>
<td>6.50%</td>
<td>1.70%</td>
<td>7.90%</td>
</tr>
<tr>
<td>Oral Hypoglycemics</td>
<td>39.20%</td>
<td>43.70%</td>
<td>44.60%</td>
<td>41.20%</td>
<td>39.70%</td>
</tr>
<tr>
<td>Proton Pump Inhibitors</td>
<td>19.70%</td>
<td>28.90%</td>
<td>31.50%</td>
<td>31%</td>
<td>30.70%</td>
</tr>
<tr>
<td>Skeletal Muscle Relaxants</td>
<td>2.90%</td>
<td>12%</td>
<td>19.50%</td>
<td>19%</td>
<td>24.80%</td>
</tr>
<tr>
<td>Statins</td>
<td>5.70%</td>
<td>11.80%</td>
<td>14%</td>
<td>14.50%</td>
<td>14.20%</td>
</tr>
<tr>
<td>Triptans</td>
<td>48.60%</td>
<td>49.60%</td>
<td>56.70%</td>
<td>58.20%</td>
<td>57.10%</td>
</tr>
<tr>
<td>Urinary Incontinence</td>
<td>40.67%</td>
<td>36.05%</td>
<td>40.07%</td>
<td>32.53%</td>
<td>33.45%</td>
</tr>
<tr>
<td>Macro-Average F-score</td>
<td>25.08%</td>
<td>31.05%</td>
<td>35.81%</td>
<td>35.54%</td>
<td>35.11%</td>
</tr>
</tbody>
</table>

*Note.* Boldface numerals on the left show the best results for TopicalMeSH vs. MeSH, whereas those on the right highlight the best performance for MeSH + words vs. TopicalMeSH + words.
Note. Boldface numerals on the left show the best results for TopicalMeSH vs. MeSH, whereas those on the right highlight the best performance for MeSH + words vs. TopicalMeSH + words.

Table 6

Comparison to tf-idf MeSH and sLDA.

<table>
<thead>
<tr>
<th>Drug Review Name</th>
<th>tf-idf MeSH</th>
<th>Topics</th>
<th>Topic+MeSH</th>
<th>TopicalMeSH</th>
<th>sLDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE Inhibitors</td>
<td>8.50%</td>
<td>24.70%</td>
<td>16.70%</td>
<td><strong>33.20%</strong></td>
<td>5.10%</td>
</tr>
<tr>
<td>ADHD</td>
<td>44.20%</td>
<td>53.40%</td>
<td>43.60%</td>
<td><strong>61.20%</strong></td>
<td>14.40%</td>
</tr>
<tr>
<td>Antihistamines</td>
<td><strong>57%</strong></td>
<td>44.60%</td>
<td>44.70%</td>
<td>40%</td>
<td>13.20%</td>
</tr>
<tr>
<td>Atypical Antipsychotics</td>
<td>37.30%</td>
<td>55%</td>
<td>47.30%</td>
<td><strong>58.20%</strong></td>
<td>43.60%</td>
</tr>
<tr>
<td>Beta Blockers</td>
<td>23.30%</td>
<td>32.80%</td>
<td>29.80%</td>
<td><strong>38.80%</strong></td>
<td>12.90%</td>
</tr>
<tr>
<td>Calcium Channel Blockers</td>
<td>29%</td>
<td>46.90%</td>
<td>39.70%</td>
<td><strong>56.80%</strong></td>
<td>29.20%</td>
</tr>
<tr>
<td>Estrogens</td>
<td>26.40%</td>
<td>43.80%</td>
<td>38.80%</td>
<td><strong>56.90%</strong></td>
<td>21%</td>
</tr>
<tr>
<td>NSAIDS</td>
<td>31.90%</td>
<td>46.40%</td>
<td>44.10%</td>
<td><strong>56%</strong></td>
<td>37.10%</td>
</tr>
<tr>
<td>Opioids</td>
<td>10.70%</td>
<td>9.30%</td>
<td>8.80%</td>
<td><strong>11.30%</strong></td>
<td>0</td>
</tr>
<tr>
<td>Oral Hypoglycemics</td>
<td>41.80%</td>
<td>49.70%</td>
<td>44%</td>
<td><strong>55.30%</strong></td>
<td>31.60%</td>
</tr>
<tr>
<td>Proton Pump Inhibitors</td>
<td>26%</td>
<td>40.30%</td>
<td>33.20%</td>
<td><strong>43.70%</strong></td>
<td>29.90%</td>
</tr>
<tr>
<td>Skeletal Muscle Relaxants</td>
<td>2%</td>
<td>9.50%</td>
<td>10.20%</td>
<td><strong>15.50%</strong></td>
<td>0</td>
</tr>
<tr>
<td>Statins</td>
<td>6.30%</td>
<td>20.30%</td>
<td>14.90%</td>
<td><strong>24.90%</strong></td>
<td>12%</td>
</tr>
<tr>
<td>Triptans</td>
<td>55.50%</td>
<td><strong>66.70%</strong></td>
<td>60.90%</td>
<td>65.20%</td>
<td>63%</td>
</tr>
<tr>
<td>Urinary Incontinence</td>
<td>39%</td>
<td>38%</td>
<td><strong>51.20%</strong></td>
<td>50.46%</td>
<td>20.2%</td>
</tr>
<tr>
<td>Macro-Average F-score</td>
<td>29.26%</td>
<td>38.76%</td>
<td>35.19%</td>
<td><strong>44.50%</strong></td>
<td>22.21%</td>
</tr>
</tbody>
</table>

Note. The best score for each corpus is shown in boldface.

In general, SVM achieved the best performance on the test data. Decision trees had the worst performance on most of these corpora, but performed well when there were a very small fraction of relevant documents. Overall, TopicalMeSH achieved a higher F-score in 14 of 15 corpora using SVMs, 12 of 15 corpora using logistic regression, and 12 of 15 corpora using decision trees. We also combined the Words features with both MeSH and
TopicalMeSH. By using SVMs or (regularized) logistic regression, TopicalMeSH + 
Words had a higher F-score than MeSH + Words in 11 of 15 corpora from both Tables 3 
and 4. However Table 5 shows that the performance of TopicalMeSH + Words and 
MeSH + Words were very close to each other using decision trees.

In Table 6, TopicalMeSH achieved a higher F-score in 12 of 15 corpora. Topics 
representation had better performance than tf-idf weighted MeSH terms and sLDA.

3.6- Section Summary and Discussion

TopicalMeSH representations outperformed those using only MeSH terms for both the 
document retrieval and classification tasks. TopicalMeSH can assess each MeSH term’s 
weight for each document and also (implicitly) assign MeSH terms to documents that 
may be missing some relevant MeSH terms. One of the challenges with making practical 
use of topic models is that users often do not understand what each topic “means” since 
the topics are distributions over words, not a single understandable word or phrase that 
summarizes the concept. TopicalMeSH effectively solves this problem by associating 
topics with corresponding MeSH terms. This provides a convenient and familiar 
terminology to present to users.

The current work has several limitations. First, the performance of LDA is highly 
dependent on the number of topics set by the user. Here I did not use a formal approach
to set this number, but rather used the number of MeSH terms as the number of topics when running LDA. This is a practical strategy that is intuitively agreeable, but determining the optimal number of topics for each corpus is challenging and may depend on the needs of the end user. For example, a smaller number of topics tends to identify more general themes, whereas larger numbers uncover more specific themes. Additional work is needed to assess how the number of topics affects the performance of the TopicalMeSH representation. Another limitation is that MeSH terms that index only a small number of documents in the corpus have very low correspondence scores to all topics. This arises because the MeSH-to-Topic correspondence matrix is based on the topic-to-Word and MeSH-to-Word distributions. Although we only consider MeSH terms indexing more than 10 documents in a corpus, there are many MeSH terms just above this threshold that have very sparse co-occurrence word distributions, leading to low correspondence scores between these MeSH terms and all topics. One possible solution to this problem is to use the full text documents, or additional documents (outside the corpus) that use these MeSH terms, to create the MeSH-to-Word distribution. Finally, all of the corpora in our evaluation were limited to systematic drug review articles, which tend to have fairly homogeneous content. Although our approach is applicable to all types of queries, we have not yet evaluated it on queries that produce more heterogeneous articles (such as free-text term searches).

For the document retrieval task, TopicalMeSH outperformed MeSH in 14 of 15 corpora. MeSH terms are created and maintained by domain experts to cover all general important
themes. TopicalMeSH combines information from human judgment and semantic information, induced by topic models, that is specific to this subset of documents. This is why TopicalMeSH has better performance in general. However, we found that MeSH’s precision-recall points were very close to TopicalMeSH’s precision-recall curves for the highly imbalanced corpora, such as Beta Blockers, Opiods, and Skeletal Muscle Relaxants. This is likely because we used only titles and abstracts to perform the evaluations, whereas MeSH is based on the entire document. The small portion of relevant abstracts may not be enough to create the relevant MeSH term’s semantic words distributions that could distinguish them from the other MeSH terms. It is also difficult for topic models to uncover latent topics to cluster a very small number of relevant articles. Hence, using those MeSH terms to find the corresponding latent topics to retrieve the relevant documents could bring in even more bias. For corpora with higher proportions of relevant articles, such as Estrogens, NSAIDS and Urinary Incontinence, TopicalMeSH performed much better than MeSH.

In the classification task, TopicalMeSH achieved a higher F-score than MeSH in general. The words feature had the worst F-score performance on these corpora using these three machine learning methods. The reason may be that the words are derived only from titles and abstracts (ignoring the MeSH information and hence information gleaned from the full texts). When I added the Words feature to both MeSH and TopicalMeSH, the latter still yielded a better performance than MeSH for most of the corpora when we used SVMs and logistic regression. In general, decision trees are not well suited to text
classification and placed too much weight on the words feature. This is why the performance of MeSH + Words and TopicalMeSH + Words in Table 5 were very similar. This demonstrates the potential of combining topic models with MeSH to induce a useful representation. TopicalMeSH is one realization of this idea. For this application, sLDA fares poorly; this may be due in part to sLDA not accounting for class imbalance.

My approach depends on having documents already indexed with MeSH terms. I use these to set both the number of topics and to choose the corresponding topic in the information retrieval task. From a practical perspective, this is not a problem for augmenting a search tool, such as PubMed with topic models to improve search results; however, this limits the applicability of our approach to corpora that are already indexed with controlled terms.
Chapter 4: Hybrid Topics

The previous chapter demonstrated that topic models can be used to address MeSH terms’ limitations, improving document retrieval and classification. However, MeSH is an important resource for PubMed and each PubMed citation has been assigned relevant MeSH terms by MeSH indexers with the help of the machine learning algorithm TMI. In this chapter, I investigate what additional utility MeSH can bring to topic models.

1) To relate MeSH terms to topics, I introduce the “Hybrid Topics”, an alternative LDA approach by modifying its original ‘bag-of-words’ assumption to ‘bag-of-MeSH&words’. By enriching each document with its indexed MeSH descriptors, ‘hybrid topics’ (mixed vectors of words and MeSH descriptors) can be generated by LDA. The objective of this chapter is to investigate whether the addition of labels (e.g. MeSH descriptors) to bags of words can facilitate the interpretation of LDA-generated topics. MeSH terms are developed to cover the most general and important themes in biomedical domains. As a result we might expect topic models of very general corpora to find general topics that are also more likely to map to existing MeSH terms. In contrast, topic models produced from a highly specialized corpus will likely reveal more specialized topics that do not map to existing MeSH terms. More specifically, to assess the interpretability of topics, I test two hypotheses using one large general biomedical corpus and one smaller specialized biomedical corpus.

1. The coherence of ‘hybrid topics’ is expected to be higher than that of regular bag-of-words topics.
2. The proportion of topics that are not associated with some MeSH descriptor is expected to be higher in a specialized corpus than in a general corpus.

The remainder of this chapter describes topic coherence, the measure used to assess the interpretation of LDA models and determine the optimal number of topics; relations between topics and MeSH terms; the test data; the evaluation experiments for the hybrid topics along with the results; and concludes with a summary and discussion of the results.

4.1- Quality of Topics

LDA is a generative model that assumes that each document is generated from a mixture of topics and that each topic corresponds to a distribution over all words in the corpus (see Chapter 3.1). The number of topics produced by LDA must be prespecified. Determining the ‘right’ number of topics for different datasets remains a challenge. When the number of topics increases, redundant and nonsense topics may be generated. Running LDA with a small number of topics will generate more general themes. In this work, I used a topic coherence measure to determine the optimal number of topics for our dataset (O'Callaghan et al., 2015).

In 2015, O'Callaghan et al. reviewed a number of topic coherence studies using various corpora and proposed a general measure based on distributional semantics, $TC-W2V$. It evaluates the relatedness of a set of top terms describing a topic, based on the similarity of their representations in a word2vec distributional semantic space. Specifically, the
coherence of a topic \( n \) represented by its top \( t \) ranked terms is given by the mean pairwise cosine similarity between all relevant term vectors in the \( \text{word2vec} \) space:

\[
Coh(t_n) = \frac{1}{\binom{t}{2}} \sum_{j=2}^{t} \sum_{i=1}^{j-1} \cos(wv_i, wv_j)
\]  

(6)

An overall score for a topic model \( M \) consisting of \( k \) topics is calculated by averaging the individual topic coherence scores:

\[
coh(M) = \frac{1}{k} \sum_{n=1}^{k} coh(t_n)
\]  

(7)

In this investigation, we use topic coherence not only to help determine the optimal number of topics, but also, more generally, the quality of topics.

4.2- Relations between MeSH and topics

Topics are generated based on the ‘bag-of-words’ assumption, which ignores word order. Each topic is represented as a list of ranked words, which is used to provide the user a sense of what this topic is about. Each document is displayed as a list of weighted topics, which represents different aspects of this document. Since the tokens within each topic are ranked according to the conditional probabilities \( P(w|t) \) learned when training the model, where \( w \) is a word and \( t \) is a topic, the top few words of each topic provide
insights into the subject of the topic. However, the interpretation of the topics (i.e., lists of words) is left as an exercise for the user.

As mentioned earlier, MeSH is developed to cover all important themes and each article in MEDLINE is indexed with a few relevant MeSH descriptors assigned by the MEDLINE indexing staff for retrieval purposes. To capture the potential relationships between MeSH and topics, we simply added the MeSH descriptors assigned to each article to the ‘bag-of-words’ for this document, creating a ‘bag-of-MeSH&word’ instead. Under this ‘bag-of-MeSH&word’ assumption, ‘hybrid topics’ are generated and each topic is represented as a list of tokens, i.e., a mixed list of ranked words and MeSH descriptors. The presence of MeSH descriptors among the top tokens for a given topic is expected to facilitate the interpretation of topics. More specifically, if a MeSH descriptor appears among the top m (for some m) tokens of a topic, we assume this MeSH descriptor is highly associated with this topic.

We consider three types of association patterns between topics and MeSH descriptors.

1) The topic has no MeSH descriptor in its top m tokens (1-0 mapping);
2) The topic has a single MeSH descriptor in its top m tokens (1-1 mapping);
3) The topic has multiple MeSH descriptors in the top m tokens (1-many mapping).

Examples of topics for each association pattern are shown in Table 7, along with the top 10 tokens for each topic.
Table 7

Topics generated based on ‘bag-of-MeSH&word’.

<table>
<thead>
<tr>
<th>Topic1</th>
<th>Topic2</th>
<th>Topic3</th>
</tr>
</thead>
<tbody>
<tr>
<td>model</td>
<td>*brain</td>
<td>motor</td>
</tr>
<tr>
<td>predict</td>
<td>cortex</td>
<td>visual</td>
</tr>
<tr>
<td>value</td>
<td>region</td>
<td>*movement</td>
</tr>
<tr>
<td>prediction</td>
<td>functional</td>
<td>*face</td>
</tr>
<tr>
<td>analysis</td>
<td>cortical</td>
<td>right</td>
</tr>
<tr>
<td>predictive</td>
<td>activity</td>
<td>response</td>
</tr>
<tr>
<td>regression</td>
<td>neural</td>
<td>*hand</td>
</tr>
<tr>
<td>datum</td>
<td>network</td>
<td>processing</td>
</tr>
<tr>
<td>estimate</td>
<td>change</td>
<td>object</td>
</tr>
<tr>
<td>predictor</td>
<td>area</td>
<td>stimuli</td>
</tr>
</tbody>
</table>

1-0 mapping  1-1 mapping  1-many mapping

Note. Asterisks indicate MeSH descriptors

4.3- Test data

One large general corpus and one small specialized corpus are used in this investigation.

The general corpus consists of 200k articles randomly selected from all PubMed articles published in 2013. The specialized corpus consists of 2472 articles from the journal Prenatal Diagnosis, which focuses on fetal medicine.

General corpus
There are about 1.2 million articles in PubMed for the year 2013. We randomly selected 200k articles (titles and abstracts) from these. This represents an appropriate amount of data given our computing resources.

To reduce the sparsity of document-to-words distribution, we performed Part of Speech tagging on the dataset and merged several categories, including NN and NNS (e.g., patient and patients); VB, VBD, VBG, and VBN (e.g., eat, ate, eaten, and eating); and JJ, JJR, and JJS (e.g., good, better, and best).

We also removed PubMed stop-words and infrequent words (with a frequency lower than 50). A total of 21,922 unique words remained. Similarly, for MeSH descriptors, I treated specific frequently used descriptors known as check tags (e.g., human, male, female, etc.) as stop words, and ignored infrequent descriptors (with a frequency lower than 5). A total of 13,853 MeSH descriptors remained.

Specialized corpus

We applied similar preprocessing to the specialized corpus, but with different cutoff values due to its smaller size. After setting a cutoff frequency of 5 for words, we obtained 3623 unique words. With a cutoff frequency of 1 for MeSH descriptors, we obtained 919 MeSH descriptors.
4.5- Evaluation Experiments

*Experiment #1*

We investigated whether the addition of MeSH descriptors to bags of words increases the quality of topics. As a surrogate for the quality of topics, I use topic coherence (Lau, Newman, and Baldwin, 2014).

To determine whether the ‘hybrid topic’ approach (i.e., ‘bag-of-MeSH&words’) outperforms the original LDA ‘bag-of-words’ approach (baseline), I generated LDA models using both approaches for a various number of topics on the two datasets. The number of topics ranged from 50 to 600 for the general corpus and from 4 to 100 for the specific corpus. For each number of topics, I calculated topic coherence for both the baseline and the hybrid topic approach.

More specifically, for the large general corpus, I used the indexed PubMed articles (titles and abstracts) published in 2013 as the background corpus when building the *word2vector* space for the original LDA with ‘bag-of-words’ assumption. To build the *word2vector* space containing both MeSH descriptors and words, I simply appended the MeSH descriptors for an article to the end of the document. In this way, I could get a mixed *word2vector* space of MeSH descriptors and words. In this experiment, I tested two different positions of MeSH descriptors in the citation (front and end) and obtained
similar topic coherence results. Following (O'Callaghan et al., 2015), I used the same
word2vec setting and the number of top terms per topic (\(t=10\)).

For the small specialized corpus, I used the full-text of these articles as the background
corpus when building the word2vec space. To build the mixed word2vector space of
MeSH descriptors and words for this background corpus, I added MeSH descriptors to
the end of each full article. In the word2vec setting for this dataset, we set vector size to
200, cutoff frequency to 3, and window size to 20.

To compare the topic coherence measures obtained within each corpus at different
numbers of topics for the baseline and the hybrid topics, I used a paired t-test.

Experiment #2

To assess whether the proportion of ‘hybrid topics’ that are not associated with some
MeSH descriptor, which reflects limited interpretability, is higher in a specialized corpus
than in a general corpus, I first have to determine the optimal number of topics in each
corpus for this assessment.

Choosing the number of topics \(k\) is a key parameter selection decision in topic modeling.
A very low \(k\) will produce overly broad topics, whereas a very high \(k\) will produce too
many small, highly similar topics. One general strategy proposed in the literature has
been to compare the topic coherence of topic models with different values of k. An appropriate value for k can then be identified by examining a plot of the mean TC-W2V coherence scores for a fixed range and selecting a value corresponding to the maximum coherence score. Since I only expected the MeSH descriptors to help interpret topics rather than for introducing new topics, I just used LDA’s original ‘bag-of-words’ assumption to determine the optimal number of topics for each test corpus.

Having determined the optimal number of topics for each corpus, I examine the ‘hybrid topics’ obtained for this number of topics and count which ones are not associated with MeSH descriptors, i.e., which ones do not contain at least one MeSH descriptor among their top-20 tokens.

I use the chi-square statistics to compare the distribution of topics of 2 patterns between the ‘hybrid topics’ and the baseline.

4.6- Results

Experiment #1

Fig. 5 and 6 display the difference in topic coherence between the ‘bag-of-MeSH&words’ assumption (hybrid topics) and LDA’s original ‘bag-of-words’ assumption (baseline) for the general and specialized corpora respectively.
Fig. 5 - Comparison of mean TC-W2V topic coherence scores for different numbers of topics $k$, generated from the general corpus
For the general corpus, we computed topic coherence for 10 different numbers of topics for both the baseline and our hybrid topics. As shown in Fig. 5, topic coherence scores are very close between the baseline and hybrid topics. The coherence is slightly better with hybrid topics after 100 topics, but slightly lower for 50 and 100 topics.

For the specialized corpus, however, a clear improvement on the coherence of topics in favor of hybrid topics compared to the baseline is displayed. As shown in Fig. 6, topic coherence scores are systematically higher for hybrid topics across all numbers of topics.

Though hybrid topics are over the baseline after 100 topics on the general corpus, the paired t-test is not significant (p=0.1624). We cannot properly assess the difference between the two approaches on this general corpus. With the specialized corpus, however, the paired t-test is highly significant (p=6.8e-25), demonstrating that the quality of the hybrid topics is better than that of the baseline topics.

Experiment #2

For the large general corpus, we generated LDA models containing $k \in [50, 600]$ topics and selected the value of $k$ that maximized mean TC-W2V coherence. As shown in Figure
7, \( k=200 \) is the first maximum and should therefore be selected as the optimal number of topics for this dataset.

**Fig. 7** - Plot of mean TC-W2V topic coherence scores for different numbers of topics \( k \), generated from the general corpus.
**Fig. 8** - Plot of mean TC-W2V topic coherence scores for different numbers of topics $k$, generated from the specialized corpus.

For the small specialized corpus, we generated LDA models containing $k \in [4, 100]$ topics. As shown in Figure 8, $k=22$ is the first maximum and should therefore be selected as the optimal number of topics for this dataset.

**Table 8**

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Optimal $K$</th>
<th># of Topic with 0 MD</th>
<th># of Topic with 1 MD</th>
<th># of Topic with $n$ MD</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Corpus</td>
<td>200</td>
<td>6 (3%)</td>
<td>16 (8%)</td>
<td>178 (89%)</td>
</tr>
<tr>
<td>Spec. Corpus</td>
<td>22</td>
<td>9 (41%)</td>
<td>6 (27%)</td>
<td>7 (32%)</td>
</tr>
</tbody>
</table>

Table 8 displays the number of different patterns of association between topics and MeSH descriptors observed in the general and specialized corpus for their respective optimal number of topics.

As shown in Table 8, the proportion of topics not associated with a MeSH descriptor is higher in the specialized corpus (41%) than in the general corpus (3%). The chi-square statistics is 57.36 with a p-value of 3.502e-13, which shows that the distributions are significantly different in these two corpora.
4.7- Chapter Summary and Discussion

This investigation demonstrates that the addition of MeSH descriptors to the traditional bag-of-words approach to creating topic models (‘hybrid topics’) can improve the quality of the topics and facilitate their interpretation, but the impact is different on a general corpus and on a specialized corpus. The quality of the hybrid topics, assessed by their coherence, is better than that of the baseline topics in the specialized corpus, but it does not seem to be the case in the general corpus.

MeSH terms are created and maintained by MeSH specialists to cover all general themes in biomedicine. However, topics extracted from a subset of documents are often specific to these documents. For the general corpus, most of the topics captured by LDA are indeed general themes. Hence, this addition of MeSH descriptors to the bag-of-words approach did not contribute much to the topic quality. This could be the reason that we did not see a significant improvement of the topic coherence score between regular topics and hybrid topics in the general corpus. In contrast, for the specialized corpus, adding MeSH descriptors can provide additional information for LDA to better differentiate between general and specific themes and to improve topic quality.

In terms of interpretability, however, the general corpus benefits from hybrid topics more than the specialized corpus does, because over 40% (9/22) of the hybrid topics remain
unlabeled (i.e., not associated with any MeSH descriptors) in the specialized corpus, compared to 3% (6/200) in the general corpus.

From the general corpus, we see that only 6 of the 200 topics (3%) contain 0 MeSH descriptors in their top 20 terms. For the specialized corpus, 9 of the 22 topics (40%) have 0 MeSH descriptors in their top 20 terms. General themes from the MeSH vocabulary may not be able to cover in detail all aspects of a specialized corpus. In contrast, the topics generated by LDA from a corpus are specific to this corpus. It is therefore logical that more topics with no MeSH descriptors are generated from a specialized corpus than a general corpus. Hence LDA will be more useful for a specialized corpus for the task of exploring concepts that may not be covered by MeSH.

From the general corpus, we also see that 178 of the 194 topics associated with MeSH descriptors (92%) are generated with multiple MeSH descriptors. MeSH descriptors are characterized in 16 top-level categories, such as category A for anatomic terms, category B for organisms, C for diseases, etc. Of these 178 topics, 140 (79%) contain MeSH descriptors from different top-level MeSH categories. These topics are most likely interdisciplinary topics. For the specialized corpus, 7 of the 13 topics associated with MeSH descriptors (54%) are generated with multiple MeSH descriptors. Topics associated with multiple MeSH descriptors from different top-level MeSH categories could be used to explore the intersection of multiple domains. LDA clearly offers an advantage for discovering interdisciplinary topics.
One limitation is that I ignored the MeSH qualifiers and only considered the MeSH descriptors when constructing our ‘hybrid topics’. In future work, I will include the qualifiers in the ‘hybrid topics’ to test whether they improve the interpretation of topic models.
Chapter 5: VizTM: A Semantic Facet Browser using Topic Models

Chapter 3 examined the power of topic models and what utilities topic model can bring in PubMed over MeSH. Chapter 4, demonstrated how adding MeSH descriptors to the “bag of words” used to create topics can improve the interpretation of topic model and uncover different types (general vs. specific) of latent themes in general vs. specific corpora. However, visualizing topic model results for general users is challenging (Blei, 2012).

Topic Browser is the visualization tool developed by (Chaney and Blei, 2012) to allow users to use topic models results to explore the large collections of documents. In our previous work (Yu, Johnson, and Kavuluru, 2013), we evaluated this Topic Browser with three biomedical researchers from different research domains. These researchers found the topic browser interesting to explore. They noted that it helped them save time reviewing the documents that they were interested in because of the topic-based presentation. One research also found it helpful for disambiguating similar topics being discussed by researchers in other related subfields. However, they also pointed out that the browser needs a way to show the documents based on a combination of topics that they are interested in. They were also more interested in starting with phrases/words instead of reviewing a single topic.

In this Chapter, I introduce an interactive visualization tool, VizTM, which is designed to allow users to easily and interactively use topic models and MeSH terms to navigate, explore, and understand large sets of biomedical documents. It provides functions to help...
users map word(s) and MeSH term(s) combinations to the most relevant concepts (topics) and then provides users the freedom to choose different topic combinations to filter the documents that they are most interested in.

This Chapter describes VizTM, and a qualitative evaluation of the tool by biomedical researchers with specific information needs.

5.1- VizTM and its Functionality

Topics are usually displayed by representing a list of top ranked words, which is almost a uniform distribution of salience over topic. Users often have difficulty using the list of words to determine the meaning of a topic. In contrast, VizTM displays each topic by using a word cloud, which uses visual features (size, density, color, and orientation) to give the user a more holistic view of each topic. Following Shneiderman’s visualization mantra—overview first, filter and zoom, then details on demand, the main overview of a set of topics consists of a single display with one word cloud per topic. Fig. 9 shows the main display of this visualization tool. The top of the page shows the PubMed query and the number of topics generated. Then there are three main
sections of this visualization tool (from left to right): tokens filter, topic overview, and documents. On the topic overview section, there are four sorting functions: sorting by ID (increasing and decreasing order); and sorting by frequency (increasing and decreasing order). The frequency of each topic refers to the number of documents related to that topic. On the documents section, there are three sorting functions under the dropdown filter: sorting by publication date (increasing and decreasing order), and sorting by relevancy. Relevancy refers to how much each document is related with the target topics selected from topic overview page.

5.1.1- VizTM - Token Filter Function
Fig. 10 displays the token filter function. Users can select a combination of words and MeSH terms to filter the relevant topics. The logical operator within words or MeSH terms is “OR”. The logical operator between words and MeSH terms is “AND”. From Fig. 10, we can see there are only three topics related with the token filter (highlighted in red box) “(datum OR model) AND (natural language processing OR data mining)”. The documents section displays the relevant documents based on the topics filtered by tokens selected. Hence, these tokens act like semantic facets. When the user selects a token, that token is first mapped to relevant concepts (topics). Only topics that contain at least one of selected words and one of selected MeSH terms in their top 200 terms will be filtered up. These topics are ranked based on the sum of the weights of the selected terms that appear in them. These topics will be used as concepts to filter out the all documents that have all of these topics in their topic distributions.
Fig. 10 - VizTM- Token Filter Function

5.1.2- VizTM- Topic Filter Function

Fig. 11 displays the topic filter function. In the figure, three topics (Topic 22, Topic 9, and Topic 37 shown in highlighted red box) were selected to filter the relevant documents. The logical operator among these three topics is “AND”, which means only documents whose topic distributions contain all three topics, will be filtered up. In this case, only 21 documents are selected given the current topics combination filter. The logical operator among topics can also be selected from the dropdown menu highlighted in the red circle in Fig. 11.
5.1.3- VizTM- Topic Page

If the user is interested in a particular topic, the user can just select that topic and navigate from the main page to the topic page shown in Fig. 12. There are three parts in this topic page (from left to right): topic detail, top related documents, and top similar topics. From the topic detail, we can see the lists of top words and MeSH terms and their corresponding weights. In the middle top related documents, there is a list of ranked documents based on the weight of this topic in each document. The right side is a list of topics represented in word could format, ranking by the similarity scores. The similarity score is calculated using cosine measure (see eq. 8) based on two topics’ words distributions:
\[
\text{Similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{8}
\]

where \(A_i\) and \(B_i\) are components of vector \(A\) and \(B\). From this page, users can select whatever document or topic they are interested in exploring.

Fig. 12- VizTM- Topic Page

5.1.4- VizTM- Document Page

User can select any document from the main page (Fig. 9) or the topic page (Fig. 12) to navigate to the document page shown in Fig. 13. There are also three parts in the document page (from left to right): top related topics, document, and top related docs.
From the left part, we can see three topics are listed using word clouds, along with a specific weight with each topic. The middle document section contains the document’s title, abstract, author information, and MeSH terms. The right section is a list of related documents ranked by the cosine similarity scores (see eq. 8) calculated based on two documents’ topic distributions. The user can navigate to any topic or document listed in this page.

Fig. 13- VizTM- Document Page

5.2- VizTM Evaluation

The aim of this evaluation was to collect initial qualitative data on whether VizTM can:
1. Provide an overview of user’s search results

2. Help users explore concepts that could better meet their information needs, such as novel concepts, interdisciplinary concepts, etc.

3. Help users explore hidden relationships between topics, documents, and documents and topics.

4. Provide a user-friendly means for users to navigate from specific word(s)/MeSH descriptor(s) to related semantic concepts (topics) and then filter relevant publications.

As we systematically evaluate the features in VizTM’s interface and develop new functionalities for PubMed search result, we conduct usability analyses to guide the development of the user interface for exploring a specific dataset. A highly usable interactive interface is essential for researchers to quickly search, explore and understand a given corpus for their use. User-centered design of information technology can improve the speed with which tasks are conducted. Therefore, we conduct usability studies (Kushniruk, Patel, and Cimino, 1997) to facilitate development of an efficient interface via an iterative process.

The study evaluates the usability of the visualization user interface using standard usability methods, including: performance measures, thinking-aloud, interviews, questionnaires, and user feedback. The resulting data was analysed using descriptive statistics as well as qualitative analysis.
5.2.1- Study Method

The usability studies involve the capture of verbal discussions and think-aloud protocols with video screen recording. In previous studies, combinations of verbal protocol analysis (Ericsson and Simon, 1985) and granular analysis of video captured data were used to characterize the relationships between actions taken and the thought processes underlying these actions (Cohen, et al., 2004; Cohen, et al., 2010).

This study (study number: HSC-SBMI-17-0388) was reviewed and approved by the Committee for the Protection of Human Subjects (CPHS) at the University of Texas Health Science Center at Houston.

Using standard usability methods, I collected data from all users recruited to this study (please see details in Study Population) and conducted quantitative and qualitative analyses. The analysed data will assist us in evaluating users’ cognitive load as well as their overall experience when interacting with the visualization tool. In addition, we anticipate that this analysis will reveal novel implicit features that are inferred during the process of exploring and understanding relevant articles. These findings will inform the development of subsequent iterations of the interface, each of which will be evaluated with respect to the relationship between ease of exploring and performance. Thus, I
evaluate the hypothesis that modifications on the basis of cognitive studies will improve user experience and system accuracy.

The study consists of two parts over two days. The first part is individually interviewing participants about their research areas and helping them create two PubMed queries given their research interests, one specific query with specific information needs, and one general query that participants are familiar with or want to explore. In Chapter 4, we evaluated the “hybrid topics” model’s performance using both a general and specialized corpus. In this study, these two queries can further help us analyze how this model helps users reach their information needs between with respect to the two different types of corpora (general corpus V.S. specialized corpus). One restriction about the general query and specific query is that PubMed results for a specific query had to contain less than 3000 citations, whereas the general query had to contain more than 3000 citations.

In the second part of the experiment, the participants were asked to use VizTM to perform tasks that we expect most users will want or need to complete, such as finding relevant articles, identifying possible meaningful topics that are novel, or identifying meaningless topics. The participants were also asked to think aloud while they were completing the tasks. They were asked to tell us, as they worked, what they were trying to do, what they were looking for, decisions that they were making, and if they were stuck or confused. After completing these tasks, they were asked to answer some questions (please see appendix A) regarding their experience with this tool.
5.2.2- Study Subjects

Participants in the study are representative of the population of biomedical researchers who have the necessary domain expertise to participate in this study. Participants had to be in the age range of 21-65 and be able to read and understand English (because the user interface is implemented in English). None of the potential candidates were excluded from the study based on gender, ethnicity, race, or physical impairments.

Recruiting was done primarily from the pool of biomedical researchers (PhD students, postdocs, faculty, etc.) in the Texas Medical Center at Houston. Prospective participants were contacted via phone and/or email. We used a standard consent form and a letter of information (see appendix B and C for details) to consent each subject. We observed each participant individually in person or remotely (via video devices) and recorded the subjects’ voice. Respondents were screened against inclusion and exclusion criteria and chosen on the basis of first-respond-first-recruit until we obtained a reasonable number of participants for our preliminary study.

In this study, we enrolled 6 participants: 2 PhD students, 2 postdoc researchers, and 2 undergraduate students working at a graduate school’s research lab. The 2 PhD students’ research areas were patient safety and neurological disorder. The 2 postdoc researchers
were working on breast cancer and liver disease. The 2 undergraduate students were working in a research lab studying dopamine.

5.2.3- Study Results

For each participant, we worked with the participant to prepare two instance of VizTM. One for a general query and the other for a specialized query. Table 9 is the details of each participant’s queries and the number of topics we used to run our “Hybrid Topic” model given each query’s results.

Table 9

Details of participants’ queries.

<table>
<thead>
<tr>
<th>Participant #</th>
<th>Query Type</th>
<th>Query</th>
<th># of Citations</th>
<th># of Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>General</td>
<td>“Patient Safety Analysis” (Last 5 Years)</td>
<td>13,366</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Specialized</td>
<td>“((medication error) OR (adverse drug event)) AND ((ontology) OR (taxonomy) OR (classification) OR (terminology))“</td>
<td>4,229</td>
<td>50</td>
</tr>
<tr>
<td>2.</td>
<td>General</td>
<td>“neurological diseases” (Last 2 Years)</td>
<td>20,754</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Specialized</td>
<td>“CSMD1” OR “RBFOX1” OR “CTNNA3” OR “MACROD2” OR “FHT” OR “CNTN4” OR “NAV2” OR “WWOX” OR “PCDH15”</td>
<td>2,071</td>
<td>50</td>
</tr>
<tr>
<td>3.</td>
<td>General</td>
<td>“hepatocellular carcinoma” (Last 4 Years)</td>
<td>24,406</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Specialized</td>
<td>“E2F and P53”</td>
<td>923</td>
<td>30</td>
</tr>
<tr>
<td>4.</td>
<td>General</td>
<td>“cell motility AND macrophage”</td>
<td>14,292</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Specialized</td>
<td>“breast cancer” AND “metastasis” AND “cell motility”</td>
<td>3,137</td>
<td>50</td>
</tr>
<tr>
<td>5.</td>
<td>General</td>
<td>“Dopamine Diseases”</td>
<td>35,025</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Specialized</td>
<td>“Dopamine Toxicity”</td>
<td>8,679</td>
<td>50</td>
</tr>
<tr>
<td>6.</td>
<td>General</td>
<td>“Dopamine Release”</td>
<td>18,572</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Specialized</td>
<td>“Drosophila Neurological Diseases”</td>
<td>336</td>
<td>15</td>
</tr>
</tbody>
</table>
The number of topics used for the specialized query is 50 in most cases. For “Drosophila Neurological Diseases”, there is only 336 citations related. Hence we used a smaller number of topics 15. For general queries, the number of return PubMed search results are over 10,000 citations. 70-topics is set for most of cases and 100-topics for those with more than 30,000 citations. To save time, we did not use topic coherence to determine the optimal number of topics for each corpus. Instead, the numbers used for each query were estimated based on our experience and the number of results in each query’s search results. We felt this was reasonable because the experiments in Chapter 4 showed that topic coherence did not vary much among topic models with different numbers of topics.

After interviewing these 6 participants for the first part our user study, we summarized their information needs as follows given the two different query types:

**General Query:**

1. Explore current hot research area;
2. Identify research area related with their interest;
3. Find studies related with their target interest (animal/human study, gene, cells, etc.);
4. Explore interactions between different areas.

**Specialized Query**

1. Explore specific concepts or themes (e.g. relations between genes, particular mutation/protein/cells, etc);
2. Find a specific publications to read.
In the second part of user study, users were first asked to use VizTM with the results of their general query. In our observation, users spent about 45 minutes on the general query to get familiar with different functions of the tool and to understand what each topic meant. When they were next presented with results of the specialized query, users showed more confidence when using the tool and spent only about 25 minutes exploring the results.

Based on our observations and post interviews, we summarize our 6 participants’ performance as following on different query types:

*Result of general query:*

1. All participants agreed that topics generated based our model can cover most aspects of their queries. These topics even show more aspects than they expected.
2. All participants could quickly find meaningful topics they would like to explore.
3. Two participants found some topics that did not make sense to them (some words were not recognized by them). However, after reviewing the relevant documents, they found those topics made sense.
4. All participants agreed that this tool could help them quickly understand and explore the general query’s results.
5. All participants identified most of topics in the general query as interdisciplinary topics.
Result of specialized query:

1. All participants could identify topics or topic combinations that met their information need in seconds.

2. All participants could find publications they wanted to read within seconds after finding target topics or topic combinations.

3. All participants agreed that VizTM could help them reach their specific information needs quickly.

4. One of the participants found topics with a small number of relevant documents less meaningful.

5. Two of the participants found multiple redundant topics in their specialized queries.

Overall, all participants found VizTM’s relation functions (document-topic, topic-topic, document-document, and term-topic relations) meaningful. They all like the topic overview function the most. They also expressed interest in continuing use VizTM in the future and were willing to recommend it to others. Two of the six participants credited VizTM with helping them to generate new hypotheses that they would like to test in their labs. Two also found their own publications and indicated that the topics assigned to their articles were appropriate. All participants did not initially realize the importance of MeSH terms. After I explained how MeSH works, they found the MeSH terms useful to help them better understand what a topic means. Based on our observations, it appeared
that topics were enough for users to identify relevant articles. Users rarely used the semantic facet functions.

The participants also suggested some improvements. They felt that topics could be displayed in different groups based on the semantic distance among topics. They would also like to have topic merge function, such that multiple redundant topics could be treated as one. Some participants suggested that phrases could be more informative than single words when representing topics. They also suggested a search function in the topic page and more metadata filters such as publication dates, publication places, etc.

5.2.3- Study Discussion

Users identified more interdisciplinary topics in the general query results and more meaningless/redundant topics in the specialized query. For PubMed search results of a general query, high level themes could be easily identified using the topic model. Hence these high level themes could be from one or more general aspects in a specific domain. This may be the reason that users found more interdisciplinary topics in the general query. For the specialized query, all topics are limited to a focused area. Some topics identified may be just from a small set of relevant citations. This small set of citations may not be enough to formalize a meaningful topic for users. It seems likely that this is why some users found topics with small numbers of relevant citations less meaningful.
5.3- Chapter Summary and Discussion

In this chapter, we introduced our interactive visualization tool of PubMed query results, VizTM, that combines topics generated using the “hybrid topics” model with an interactive visualization tool for navigating words, MeSH terms, topics, and documents. A preliminary user study was conducted with six participants. They all found that the tool could help them quickly understand PubMed search results and identify interesting articles that they would like to read. One limitation of this work is that we only included 6 participants in our study, which may not be representative for Texas Medical Center’s biomedical researchers. In the future, more participants will be recruited in the study and more statistical analysis will be conducted to better support our results.
Chapter 6: Conclusion, Limitation, and Future Work

6.1- Conclusion

As the amount of unstructured information grows in biomedicine, researchers will need more tools to overcome information overload. Although tools that quickly identify relevant information are helpful and still needed, new tools are needed to help users quickly gain an understanding of large sets of relevant sources, and to identify latent themes and relationships. This dissertation addresses these needs in three different, but connected ways. First, it introduced a method for relating discrete human curated tags (MeSH descriptors) with continuous semantic vector representations (topics) to derive a combined semantic representation (TopicalMeSH) for PubMed citations. This is the first evaluation of the correspondence relationship between MeSH terms and topic models. We used 15 systematic drug review corpora to evaluate performance on information retrieval and classification tasks using the TopicalMeSH representation, compared to using standard encodings that rely on either (1) the original MeSH terms, (2) the text, or (3) their combination. For the document retrieval task, we compared the precision and recall achieved by ranking citations using MeSH and TopicalMeSH representations, respectively. This proposed TopicalMeSH representation (which combines MeSH terms with latent topics) consistently improved performance on document retrieval and classification tasks, compared to using alternative standard representations using MeSH terms alone, as well as, several standard alternative approaches.
Second, this work investigated whether the addition of labels (e.g., MeSH descriptors) to bags of words can improve the quality and facilitate the interpretation of LDA-generated topics. We added MeSH descriptors to the ‘bag-of-words’ assumption to generate ‘hybrid topics’, which are mixed vectors of words and descriptors. We evaluated this approach on the quality and interpretability of topics in both a general corpus and a specialized corpus. Our results demonstrated that the coherence of ‘hybrid topics’ is higher than that of regular bag-of-words topics in both the specialized corpus and the general corpus. We also found that the proportion of topics that are not associated with MeSH descriptors is higher in the specialized corpus than in the general corpus. The specific contribution of this work is to introduce an alternative LDA approach by changing its original ‘bag-of-words’ to a ‘bag-of-MeSH&words’ approach. By enriching each document with its indexed MeSH descriptors, ‘hybrid topics’ (mixed vectors of words and MeSH descriptors) can be generated by LDA. Topics associated with multiple MeSH descriptors from different top-level MeSH categories could be used to explore the intersection of multiple domains. LDA clearly offers an advantage for discovering interdisciplinary topics.

The last part of this dissertation introduced a visualization tool (VizTM) based on the “hybrid topics” model to allow users to interactively use topic models and MeSH terms to understand and explore PubMed search results. Six biomedical researchers from the Texas Medical Center at Houston participated in a user study of this tool. Though they
each had different information needs, all of them found that VizTM could help them quickly understand search results for both general queries and specialized queries, as well as identify interesting topics and publications. VizTM provides an alternative view of PubMed search results. Instead of presenting results as a linear list of items over many pages, VizTM provides an overview of different concepts automatically summarized by a topic model. With this overview, users can easily identify whether the results are appropriate and find interesting areas they would love to explore. Preliminary results suggest that this tool is useful for biomedical researchers when exploring new research domains, keeping updated with current research domains, or finding specific information from large volumes of search results.

6.2- Limitations

While this work showed clear improvements of TopicalMeSH over MeSH, topic models alone, and other alternative approaches, the TopicalMeSH approach is limited to documents already indexed with MeSH terms or other human curated tags. Another limitation is that the performance of LDA is dependent on the number of topics set by the user. Here we did not use a formal approach to set this number, but rather used the number of MeSH terms as the number of topics when running LDA. This is a practical strategy that is intuitively agreeable, but determining the optimal number of topics for each corpus is challenging, may depend on the needs of the end user, and can greatly affect the utility of the resulting topics.
In the second part of this dissertation, we ignored the MeSH qualifiers and only considered the MeSH descriptors when constructing ‘hybrid topics’. MeSH qualifiers are an important part to help represent different aspects of MeSH descriptors. Without considering MeSH qualifiers, we may misinterpret some MeSH descriptors. We also used only one domain “Prenatal Diagnosis” as the specialized corpus in our evaluation dataset. Though “hybrid topics” performs better than regular topics in this specialized corpus, we did not consider whether that performance improvement may transfer to other domains.

Though our visualization tool, VizTM, could help users quickly understand and explore large PubMed search results, it cannot display results immediately after the user submits their queries. Constructing topic models is a time consuming, computationally expensive process. Although distributed LDA (Xing, al et., 2015) can help dramatically reduce topic model construction time to minutes, it still falls short of real-time interactive performance. The evaluation of VizTM is also limited because we estimated a robust number of topics given the number of citations of each query’s result. This estimate may not be the optimal number for users to explore PubMed search results. The study is also very preliminary and qualitative with only have six participants, which may not be representative for the whole biomedical researchers in Texas Medical Center at Houston. Finally, we did not compare VizTM with other tools such as PubMed.

6.3- Future Work
MeSH (Medical Subject Headings) provides a hierarchically-organized terminology for indexing and cataloging biomedical information such as MEDLINE/PubMed and other NLM databases. In our work, we treated each MeSH term independently. In the future, we plan to add this additional structural information into the topic model results. In our tool, we used topic models to help map a term to a concept to filter relevant documents. In the future, we’d like to continue explore other semantic distribution methods (Niu, et al., 2015, Wang & Koopman, 2017) to support the semantic facet function. We will also explore other distributed computational methods for constructing topic models (Newman, et al., 2009, Li, Kluger, & Tygert, 2016) to improve our tool’s users experience.
References


topic model and domain knowledge. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on* (pp. 191-192). IEEE.


Sarkar, I. N., Schenk, R., Miller, H., & Norton, C. N. (2009, November). LigerCat: using “MeSH clouds” from journal, article, or gene citations to facilitate the
identification of relevant biomedical literature. *American Medical Informatics Association.*


Appendix A: User Study Questions

Semantic Facet Browser using Topic Model User Testing

Thank you so much for taking time to fill out the form. Your feedback helps us to improve the usability of our visualization tool.
Please enter your name:

General query part
1- Can those topics cover most aspects of this general query? If no, please list those aspects that this tool doesn’t mention.

2- Do you find any meaningful topics that you are not expected? If so, please list them

3- Do you find any meaningful topics that may not be described by MeSH terms? If so, please list them.
4- Do you find any meaningless topics? If so, please list them.

5- Do you find any interdisciplinary topics? If so, please list them.

6- Do you find this tool useful to help you understand this general query’s result?

**Specific query part**
1- Do you find any topics or topic combinations that meet your specific information needs? If so, please list them.

2- Do you find any good articles to meet your information needs?

3- Do you find any meaningful topics that may not be described by MeSH terms? If so, please list them.

4- Do you find any meaningless topics? If so, please list them.

5- Do you find any interdisciplinary topics? If so, please list them.

6- Do you find this tool useful to help you reach your specific information needs?

**Overall**
1- Is it obvious to do what you need using this tool? If no explain.

2- Can you use this tool without having to read a manual?

3- Does token filter function helpful to select relevant topics you interested in?

4- Does the topic combination filters helpful to filter relevant documents

5- Does the similar topics function helpful to find relevant topics?

6- Does the similar documents function helpful to find relevant documents?

7- Is there any metadata you think would be helpful for your information needs?

8- Can you find the facets easily?

9- Do you understand what facets do?

10- What are the additional functions that you would like to see in this tool?

11- Was there something missing, you were expecting to see?
12- Was anything too obtrusive?

13- What would encourage you to return to this site in the future?

14- Name your three favorite things about the site, and your three least favorite.

15- If you could change one thing on the site, whether it is major or minor, what would be at the top of the to do list?

16- How well is the user interface structured?

17- Do you think there is too much information on the screen?

18- What is your opinion of using word cloud to represent a topic?

19- What part of the interface stands out the most?

20- Is the data on the website easy to read (both font style and size)?

21- How intuitive and helpful is the navigation system?

22- Will you recommend this tool to others?
Appendix B: Information Consent

You are invited to take part in a research project called, Usability Analysis for the Semantic Facet Browser using Topic Models for PubMed Search Results, conducted by Dr. Todd Johnson, of the University of Texas Health Science Center. For this research project, he will be called the Principal Investigator or PI.
Your decision to take part is voluntary. You may refuse to take part or choose to stop from taking part, at any time.
You may refuse to answer any questions asked or written on any forms. This research project has been reviewed by the Committee for the Protection of Human Subjects (CPHS) of the University of Texas Health Science Center at Houston as HSC-SBMI-17-0388.

PURPOSE

The purpose of this research study is to evaluate the usability of the visualization tool developed to help users more efficiently and effectively explore and understand PubMed search results.
The study will be conducted in person and online using web-based teleconferencing tools. The target audiences are among stakeholders who wish to essentially assist us in developing this data exploration tool. This is a local study conducted in Houston TX. The study will enroll a total of 10 people worldwide. This location will enroll 10 people.

PROCEDURES

If you agree and are able to take part in this study you will perform the following activities:
Use a computer to access this visualization tool.
Interact with the tool by exploring its functions while also paying attention to its cosmetic features. For example, use topics to help summarize PubMed search results while also taking the color code used for displaying topics into consideration. Think aloud while you are performing any activities on the tool. For example, you need to voice aloud what you are trying to do, what you are looking for, decisions that you are making, or if you are stuck or confused. Basically verbalizing your thought as you move through the interface. We will observe and take notes about how you use the tool and any comments that you make regarding the tool’s usability, usefulness to you, your own search strategies, and so on. Your interactions with the system as well as your voice will be recorded using standard screen recording software.

**TIME COMMITMENT**

The total amount of time you will take part in this research study is up to 1 hour.

**BENEFITS**

You may receive no direct benefit from being in the study; however, The University of Texas Health Science Center at Houston may benefit from your participation and/or what is learned in this study.

**RISKS AND/OR DISCOMFORTS**

The anticipated risk of this study is no greater than the risk of everyday activities. However, the study may include risks that are unknown at this time. Confidentiality: Possible risk of breach of confidentiality

**Questionnaires:** You may get tired when we are asking you questions or you are completing questionnaires. You do not have to answer any questions you do not want to answer.

**ALTERNATIVES**

The only alternative is not to take part in this study.

**STUDY WITHDRAWAL**

Your decision to take part is voluntary. You may decide to stop taking part in the study at any time.
The PI or test proctor can stop the study at any time if the domain expertise is not appropriate or sufficient for performing the annotation of the provided material.

The information collected until the termination will be used for further analyses in the study.

COSTS, REIMBURSEMENT AND COMPENSATION

If you decide to take part in this research study, you will not incur any costs. You will not be paid to take part in the study.

CONFIDENTIALITY

You will not be personally identified in any reports or publications that may result from this study. Any personal information about you that is gathered during this study will remain confidential to every extent of the law. A special number (code) will be used to identify you in the study and only the investigator will know your name.

NEW INFORMATION

If you are interested in the final results, please provide your name and email address independently to the study team. We will be happy to share the aggregate results and associated publication(s) with you.

QUESTIONS

If you have questions at any time about this research study, please feel free to contact Todd Johnson at 713-500-3913 (or todd.r.johnson@uth.tmc.edu, as they will be glad to answer your questions. You can contact the study team to discuss problems, voice concerns, obtain information, and offer input in addition to asking questions about the research.

CPHS STATEMENT: This study (HSC-SBMI-17-0388) has been reviewed by the Committee for the Protection of Human Subjects (CPHS) of the University of Texas Health Science Center at Houston. For any questions about research subject's rights, or to report a research-related injury, call the CPHS at (713) 500-7943.
INFORMED CONSENT FORM TO TAKE PART IN RESEARCH

Title: “Usability Analysis for the Semantic Facet Browser using Topic Model for PubMed Search Result”

Letter of Information
(HSC-SBMI-17-0388)
Primary Investigator: (Dr. Todd Johnson)

You are invited to take part in a research study called, “Usability Analysis for the Semantic Facet Browser using Topic Model for PubMed Search Result”, conducted by Dr. Todd Johnson, professor of the University of Texas Health Science Center at Houston. For this research project, he will be called the Principal Investigator or PI.

The purpose of study is to see if evaluate the usability of the visualization tool developed to help users more efficiently and effectively explore and understand PubMed search results. If you decide to take part in the study the total time commitment is 1.5 hours. You are invited to take part in this study because you are biology/biomedicine/biomedical researcher. Our tool is developed to help you understand large collection of publications more easily and quickly. You can refuse to answer any questions asked or written on any forms. Participation in this study is voluntary. A decision not to take part in this study will not change the services or your employment.

If you agree to take part in this survey you will agree to about a 30-minutes survey.

You may not receive any benefit from taking part in this study. The information you provide will help us better develops this next generation of documents exploration tool. The only possible risk may be breach of confidentiality. This information collected will not contain identifying information. You have the alternative to choose to not take part in this study and can withdraw at any time.

There is no cost and you will not be paid to take part in this study. You will not be personally identified in any reports or publications that may result from this study. Any
personal information about you that is gathered during this study will remain confidential to every extent of the law.

If you have any questions about this project please contact (Dr. Todd Johnson, phone: 713.500.3913)

If you agree to take part in the study your agreement is completion of the survey or you will remain present for focus group discussion.

This research project has been reviewed by the Committee for the Protection of Human Subjects (CPHS) of the University of Texas Health Science Center at Houston (HSC-SBMI-17-0388)
For any questions about research subjects’ rights, please call CPHS at (713) 500-7943.