



University of Kentucky
UKnowledge

Theses and Dissertations--Electrical and
Computer Engineering

Electrical and Computer Engineering

2013

Cooperative Semantic Information Processing for Literature- Based Biomedical Knowledge Discovery

Zhiguo Yu
University of Kentucky, zhiguoyu1989@gmail.com

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Yu, Zhiguo, "Cooperative Semantic Information Processing for Literature-Based Biomedical Knowledge Discovery" (2013). *Theses and Dissertations--Electrical and Computer Engineering*. 33.
https://uknowledge.uky.edu/ece_etds/33

This Master's Thesis is brought to you for free and open access by the Electrical and Computer Engineering at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Electrical and Computer Engineering by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained and attached hereto needed written permission statements(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine).

I hereby grant to The University of Kentucky and its agents the non-exclusive license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless a preapproved embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's dissertation including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Zhiguo Yu, Student

Todd R Johnson, Major Professor

Zhi David Chen, Director of Graduate Studies

COOPERATIVE SEMANTIC INFORMATION PROCESSING FOR LITERATURE-BASED BIOMEDICAL KNOWLEDGE DISCOVERY

THESIS

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering in the College of Engineering at the University of Kentucky

By

Zhiguo Yu

Lexington, Kentucky

Directors: Dr. Todd Johnson, Professor of Biostatistics Department

Lexington, Kentucky

2013

Copyright © Zhiguo Yu 2013

ABSTRACT OF THESIS

COOPERATIVE SEMANTIC INFORMATION PROCESSING FOR LITERATURE-BASED BIOMEDICAL KNOWLEDGE DISCOVERY

Given that data is increasing exponentially everyday, extracting and understanding the information, themes and relationships from large collections of documents is more and more important to researchers in many areas. In this paper, we present a cooperative semantic information processing system to help biomedical researchers understand and discover knowledge in large numbers of titles and abstracts from PubMed query results.

Our system is based on a prevalent technique, topic modeling, which is an unsupervised machine learning approach for discovering the set of semantic themes in a large set of documents. In addition, we apply a natural language processing technique to transform the “bag-of-words” assumption of topic models to the “bag-of-important-phrases” assumption and build an interactive visualization tool using a modified, open-source, Topic Browser. In the end, we conduct two experiments to evaluate the approach. The first, evaluates whether the “bag-of-important-phrases” approach is better at identifying semantic themes than the standard “bag-of-words” approach. This is an empirical study in which human subjects evaluate the quality of the resulting topics using a standard “word intrusion test” to determine whether subjects can identify a word (or phrase) that does not belong in the topic. The second is a qualitative empirical study to evaluate how well the system helps biomedical researchers explore a set of documents to discover previously hidden semantic themes and connections. The methodology for this study has been successfully used to evaluate other knowledge-discovery tools in biomedicine.

KEYWORDS: Data Mining, Topic Modeling, Knowledge Discovery, Natural Language Processing, Information Visualization

Zhiguo Yu

April 14, 2013

COOPERRATIVE SEMANTIC INFORMATION PROCESSING FOR LITERA-
TURE-BASED BIOMEDICAL KNOWLEDGE DISCOVERY

By

Zhiguo Yu

Dr. Todd Johnson

Director of Thesis

Dr. Zhi Chen

Director of Graduate Studies

April 14, 2013

ACKNOWLEDGMENTS

This publication was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, US National Institutes of Health (NIH), through Grant UL1TR000117. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

I would like to thank my academic advisor, Dr. Todd R Johnson, for the opportunity he gave me to pursue my Master degree, and all the guidance and help I've received from him all through these years. This thesis would be impossible without his extensive knowledge and innovative ideas in this field.

Special thanks should be accorded to Dr. Ramakanth Kavuluru for providing technical assistance at the Natural Language Processing group and the insightful guidance for the evaluations of both the models and the interface.

Thanks to my labmate, Sifei Han, for the time and effort she spent on the C-value method.

Last but not least, I would like to express my deepest gratitude to my parents and my brother, for the endless love and support I have always been with since I was born.

TABLE OF CONTENTS

ACKNOWLEDGMENTS.....	iii
LIST OF TABLES.....	v
LIST OF FIGURES	vi
Chapter 1 Introduction.....	1
1.1 Knowledge Discovery	1
1.2 Topic Models.....	2
1.3 Motivation.....	3
Chapter 2 Background.....	7
2.1 Generative Model	7
2.2 Latent Dirichlet Allocation (LDA)	9
2.3 Topical n-gram Model.....	13
2.4 C-value Method	15
2.5 Topic Browser	18
Chapter 3 Methods.....	21
3.1 Phrase LDA method	21
3.2 Evaluation of Models.....	21
3.3 Phrase/n-gram Intrusion Test.....	25
3.4 Evaluation of Topic Browser	27
Chapter 4 Results	31
4.1 Results of Models' evaluation	31
4.2 Results of Topic Browser's evaluation.....	31
Chapter 5 Conclusion and Future Work.....	37
5.1 Conclusion	37
5.2 Future Work.....	37
Bibliography	39
Vita.....	43

LIST OF TABLES

Table 1: Notation used in this paper	10
Table 2: Example for C-value method.....	17
Table 3: Three sample topics generated by LDA, topical n-gram model, and Phrase LDA. Top ten words or phrase are listed for each topic.	24
Table 4: Three evaluations of the Topic Browser.....	34

LIST OF FIGURES

Figure 1: Nine sample topics generated by LDA. Each list is a topic and represented by top 10 words ordered by the conditional probability $p(w/t)$, where w is the word and t is the topic.....	6
Figure 2: Sampling from a single coin.....	9
Figure 3: Sampling with repeated choice of coin	9
Figure 4: Graphical Model of LDA	12
Figure 5: Graphical model of the topical n-gram Model	14
Figure 6: Topics summarization page of the Topic Browser	19
Figure 7: Topic page of the Topic browser.....	20
Figure 8: Document page of the Topic Browser.....	20
Figure 9: Comparison of Word and Phrases selected Distributions	23
Figure 10: Example of the questionnaire. Each question has 4 choices and only one of them is the intruder.	30
Figure 11: Topics summarization page for subject 1	35
Figure 12: Topics summarization page for subject 2.....	35
Figure 13: Topics summarization page for subject 3.....	36

Chapter 1 Introduction

1.1 Knowledge Discovery

Knowledge discovery is a fundamental and important activity in biomedical research. We all acknowledge the fact that knowledge is the end product of a data-driven discovery process [17]. Extracting and understanding the knowledge, information, themes and relationships from large collections of documents is an important task for biomedical researchers. It's common for a biomedical researcher to read and analyze published articles in his or her area of expertise to get targeted information or just to stay up-to-date. However, it is getting increasingly difficult for researchers to keep up with even narrowly defined research areas. The approximate number of published paper by the end of 2008 was 49,234,626, which grew to 50,712,009 by the end of 2009 [1]. PubMed as one of the most popular database on biomedical and life science topics has more than 22.6 million records as of today. For example, there will be 195,106 papers associated with the “diagnostic imaging” and “cardiovascular system” or 159,661 if limited to human studies, core clinical journals, and Medline. If a specialist wants to read all of these papers, it will take 11 years and 124 days with 8 hours a day, 5 days a week and 50 weeks a year. By the same time, there will at least 82,142 more papers added and these will take another 8 years and 78 days to finish [2].

The growth in the amount of existing data has exceeded the limit of human's ability to manually read, understand, and organize. . Specialization in narrow areas is no longer sufficient to tame the problem. In addition, the problem is growing worse because of the increasing need for researchers to work across different areas and traditional scien-

tific silos. Translational, multidisciplinary, and multilevel research requires researchers to understand and find links across disparate bodies of knowledge. In 1996 Usama Fayyad said, "... There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data...." [3]. This thesis explores one such tool, topic models, to help humans explore and understand large numbers of publications.

1.2 Topic Models

Topic Models are a family of unsupervised learning algorithms developed to discover the hidden semantic structure of a large collection of documents. A topic model and corresponding user interface for browsing the model can help researchers search, browse, summarize and organize large archives of text documents [7]. Topic models extract a set of semantic themes from large electronic archives and assign multiple themes to each document. These themes are called "topics." Once the topics are extracted, they can be used for classification, summarization, information retrieval, visualization and so on. The goal of topic models is to produce an efficient and convenient way to discover new information or knowledge from large sets of documents [8].

Latent Dirichlet Allocation (LDA) [4], one of the most popular topic modeling methods, is an unsupervised machine learning approach that can be viewed as a three-level hierarchical Bayesian model. It has already been applied in the context of biomedical research, for example, in the psychology domain for predicting behavior codes arising from couple therapy transcripts [5] and for risk stratification in ICU patients [6] (Lehman

Lw) using nursing text from the first 24-hours of patients' ICU stays.

Unlike clustering approaches where documents are grouped into mutually exclusive clusters based on document-based features, topic models represent each document as a mixture of different topics, each topic as a distribution of unique words with the topics varying in probability across all documents. Finally the topics are represented as bags of words where only the top m words (for some m) are shown for each topic. Since the words within each topic are ranked according to the conditional probabilities $p(w/t)$ learned when training the model, where w is the word under topic z , the top few words of each topic provide insights into the subject of the topic. Figure 1 is an example of topics generated by LDA based on 26,533 documents fetched from PubMed with the query ‘*public health [majr] AND united states [mh] AND “last 4 years”*’

1.3 Motivation

The original LDA model was developed based on the popular “bag-of-words” assumption, in which the word order and phrases are ignored. In many applications, results of LDA were found to contain ambiguous lists of words as representatives of the topics because of the inherent polysemy and homonymy of words. As a result, researchers may have trouble understanding what a topic is about and how some topics differ.

In general, single words convey less information than phrases. Some verbs or prepositions are even meaningless without related words. For example, the meaning of “magnetic resonance imaging” cannot be completely determined from any one of these

three words in isolation, “magnetic”, “resonance” or “imaging”. Thus the “bag-of-words” assumption can not always meet the needs of extracting salient themes from large sets of documents. In 2006, Wallach developed a bigram topic model [9] based on the original LDA (or just LDA), in which she incorporates bigram statistics into the latent topic variables to add the dependencies between consecutive words. In 2007, Wang et al. presented another topic model, called the topical n-gram model [11], based on Wallach's bigram model, which can form longer n-grams for $n > 2$. Even though the topical n-gram model approach enriches the generated topics by longer sequences of words, the topic generation process is still based on individual words with the word context providing evidence to form a longer n-gram. We call this approach the “bag-of-n-grams” method.

In this paper, we propose a new LDA based model called the Phrase LDA where the topics are generated based on “important” noun phrases instead of words or n-grams; thus our approach can be called the “bag-of-key-phrases” approach. We use the C-value method [12] for extracting the key phrases and build the LDA model based on the key phrases that have a C-value score (more on this later) that is above a certain threshold. A user study with 11 participants using the “word intrusion” test [13] for topic model evaluation demonstrates that the Phrase LDA approach provides 7% improvement over the topical n-gram model. 8 out of 11 participants also answered that it was easy to comprehend the phrase LDA models.

Given that topic models are high-level tools to summarize the corpus, the outputs of topic models are not easy to understand by users who are not familiar with these

models and numerical distributions [15]. Hence, an efficient, effective and convenient way is needed to interact and visualize the topics, documents and corpus. In 2012, Chaney and Blei designed a visualization tool, called the topic browser [14], to present the summarization of the corpus, reveal the relationships between document and topics and the relationship between documents. In our system, we applied a modified topic browser to our cooperative semantic information processing for literature based biomedical knowledge discovery system. We evaluated how well the system helps biomedical researchers explore a set of documents to discover previously hidden semantic themes and connections. Information visualizations are difficult to evaluate, because they are primarily tools for supporting a creative process for developing insight and generating and then exploring hypotheses using open-ended discovery [16]. Thus a key measure of success of visualizations is whether they help biomedical researchers develop new questions and new hypotheses, not to simply answer pre-existing questions. We used the qualitative evaluation methodology developed by Saralya, North, and Duca for evaluating how well microarray visualization tools enabled biological insight.

Topic 1	Topic 2	Topic 3
quality	clinical	cost
medical	trials	costs
electronic	trial	per
data	studies	life
information	treatment	effectiveness
records	randomized	economic
health	study	model
record	research	million
care	results	treatment
patient	controlled	benefits
Topic 4	Topic 5	Topic 6
scores	drug	care
validity	fda	health
measures	products	services
scale	food	medical
reliability	drugs	patients
quality	administration	visits
factor	devices	primary
measure	device	medicare
items	product	emergency
item	safety	insurance
Topic 7	Topic 8	Topic 9
radiation	species	data
dose	control	information
sleep	mosquito	monitoring
noise	traps	national
hearing	wnv	based
imaging	vector	collection
exposure	host	used
doses	deer	analysis
image	tick	methods
effective	field	network

Figure 1: Nine sample topics generated by LDA. Each list is a topic and represented by top 10 words ordered by the conditionally probability $p(w/t)$, where w is the word and t is the topic.

Chapter 2 Background

In this section, we provide a brief background on the original LDA, topical n-gram model, the C-value method for key phrase extraction and visualization tool, topic browser.

2.1 Generative Model

In probability and statistics, a generative model describes a process, usually one by which observable data is generated given some hidden parameters. In the simplest case, the model generates samples independently, which means there is no dependency between any two random samples [20]. Let's take a coin as the generative source. If it is a fair coin, flipping it will result in a tail or head based on a uniform distribution: [19]

$$f(i) = \frac{1}{2} \quad \text{for } i \text{ in } \{ \text{tail}, \text{head} \}$$

Hence, the probability of a particular coin sequence can then be calculated by the product of the probability of individual observation.

$$p\{i_1, i_2, \dots, i_k\} = \prod_{j=1}^k p(i_j)$$

If we treat a tail as 0 and a head as 1, this model can be viewed as a source that can generate numbers 0 or 1 according to the uniform distribution.

Generative models can also have a hierarchical structure. For example, we have two coins. One is a fair coin and the other is a trick (biased coin):

$$\text{If the coin A is fair, } f(i) = \begin{cases} \frac{1}{2} & , \quad i = \textit{tail} \\ \frac{1}{2} & , \quad i = \textit{head} \end{cases}$$

$$\text{If the coin B is the trick one, } f(i) = \begin{cases} \frac{4}{5} & , \quad i = \textit{tail} \\ \frac{1}{5} & , \quad i = \textit{head} \end{cases}$$

Now let's imagine the following two random processes:

1. Pick a coin from these two coins according to a uniform distribution;
2. Flip the chosen coin.

Based on this generative process, the probability of one sample should be:

$$P(i) = P(A) \cdot P(i / A) + P(B) \cdot P(i / B) = \begin{cases} \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{4}{5} = \frac{13}{20} & i = \textit{tail} \\ \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{5} = \frac{7}{20} & i = \textit{head} \end{cases}$$

This model is hierarchical because the final outcome of each flip depends on the probability of choosing between the two coins. Models like this are also called mixture models because each observation depends on a mixture of several random choices. The weight of each observation is a sum of different distributions. In addition, the distributions in these processes can be any distribution we want. Mixture models are widely used in modern probabilistic modeling because they permit probabilistic reasoning and analysis of phenomena using complex, interdependent representational structures.

Figures 2 and 3 are graphical representation of generative models [21]. Nodes are random variables. Solid nodes are observable variables and empty nodes are unob-

servable (latent) variables. The edges show the dependency between nodes. The plates around nodes indicate the repetitions. Figure 2 means that we use one coin for the whole process and then repeatedly flip it N times. Figure 3 represents that for each of the N times, we pick a new coin first and then flip it.

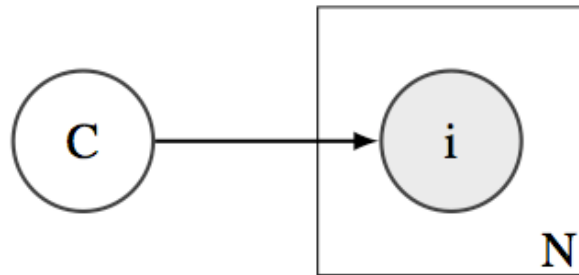


Figure 2: Sampling from a single coin

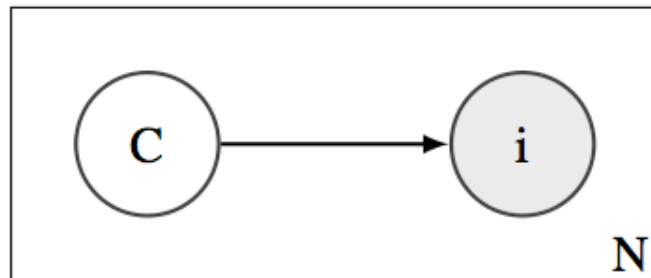


Figure 3: Sampling with repeated choice of coin

2.2 Latent Dirichlet Allocation (LDA)

Topic Models are a very popular class of mixture-based models. They have been applied to document classification, clustering and information retrieval. Topic models treat

each document as “bag-of-words” and assume that the words in each document exchangeable [4], which means the joint probability of words in each document is

Table 1: Notation used in this paper

Symbol	Description
w_n	Nth word in document
D	Documents in corpus
d	Single document
z	Topic
ϕ_z	Topic to words distribution for topic z
β	Dirichlet prior for ϕ
α	Dirichlet prior for θ
θ_d	Document to topics distribution for document d
δ	Dirichlet prior for σ
$\sigma_{z,w}$	Bigram distribution for each word w in topic z
γ	Beta prior for ψ
$\psi_{z,w}$	Bigram status for topic z and word w
x_n	Bigram status for the nth word
ϕ_{z_d}	Topic z in document d to words distribution

invariant to any permutation of these words. If we assume π as a permutation of the integers from 1 to N:

$$p(w_1, \dots, w_N) = p(w_{\pi(1)}, \dots, w_{\pi(N)}),$$

where w is the word in a document.

In 2003, Blei, Ng and Jordan developed Latent Dirichlet Allocation (LDA), which is perhaps the most well known “bag-of-words” topic model. In the LDA model, a document is represented as a mixture of latent topics and each topic is represented as a distribution of unique words. In the generative modeling perspective, LDA represents a corpus of documents at three levels: the corpus level, the document level, and the word level as follows:

1. At the corpus level, LDA generates a topic-words distribution ϕ_z for each topic z from the topic-words Dirichlet prior β ;
2. At the document level, LDA generates a document-topics distribution θ_d for each document d from the document-topics Dirichlet prior α ;
3. At the word level, LDA generates the topic assignment z_n from the document-topics distribution θ_d first and then generates a word assignment from the topic-words distribution ϕ_{z_d} for each word w_n in document d .

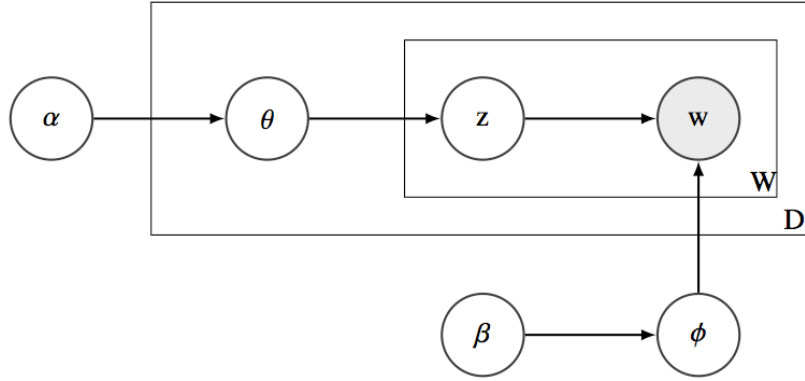


Figure 4: Graphical Model of LDA

Figure 4 is a graphical model representation of LDA. The α and β are Dirichlet priors as explained in the list above at the corpus level and document level. D and W plates in this figure consist of distributions at the document level and word level respectively. The joint probability of this generative process is:

$$p(\theta, \phi, z, w | \alpha, \beta) = \left(\prod_{t=1}^T p(\phi_t | \beta) \right) \left(\prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | z_{d,n}, \phi) \right)$$

where T is the number of topics, D is the number of document in corpus and N is the number of words in a particular document. From Figure 4, we can see only the word w in each document is observable. Hence the central inference problem is to define the posterior probability from the joint probability.

$$p(\theta, \phi, z | w, \alpha, \beta) = \frac{p(\theta, \phi, z, w | \alpha, \beta)}{p(w | \alpha, \beta)}$$

However, this posterior distribution is intractable to compute in general. Hence some approximated posterior inference algorithms have been developed [26] [27]. For example, mean field vibrational methods [22] [23], expectation propagation [24], collapsed Gibbs sampling [25], collapsed variational inference [28], online variational inference [29], Markov chain Monte Carlo sampling [30], and optimization-based variational inference [4], which is used in our system.

2.3 Topical n-gram Model (TNG)

This original LDA approach is based on the bag-of-words approach, where the words w are conditionally independent given their assigned topic z . However, as discussed in chapter 1, the word-based topics are often not informative. This leads to the development of the topical n-gram model [11], where two more dependencies are introduced at the word level. The first is the dependency between two consecutive words, the other is the dependency on the bigram status, which determines whether a bigram needs to be formed for the same consecutive word tokens depending on their nearby context. This model can also be expressed at three levels:

1. At the corpus level
 - a) TNG generates a topic-words distribution ϕ_z for each topic from the topic-words Dirichlet prior β ;
 - b) TNG generates the bigram status Bernoulli distribution $\psi_{z,w}$ for each topic z and each word w from the Beta prior γ ;
 - c) TNG generates the bigram distribution $\sigma_{z,w}$ for each topic z and each word w from the Dirichlet prior δ

2. At the document level

- a) TNG generates a document-topics distribution θ_d for each document d in the corpus from the document-topics Dirichlet prior α ;

3. At the word level

- a) TNG generates a topic assignment z_n from the document-topics multinomial distribution θ_d ;
- b) TNG generates a bigram status x_n for each word w_n in document d from the Bernoulli distribution $\psi_{z_{n-1}, w_{n-1}}$;
- c) If the bigram status $x_n = 1$, TNG generates the word assignment w_n from the bigram distribution $\sigma_{z_n, w_{n-1}}$, else, TNG generates the word assignment w_n from the topic-words distribution ϕ_{z_n} .

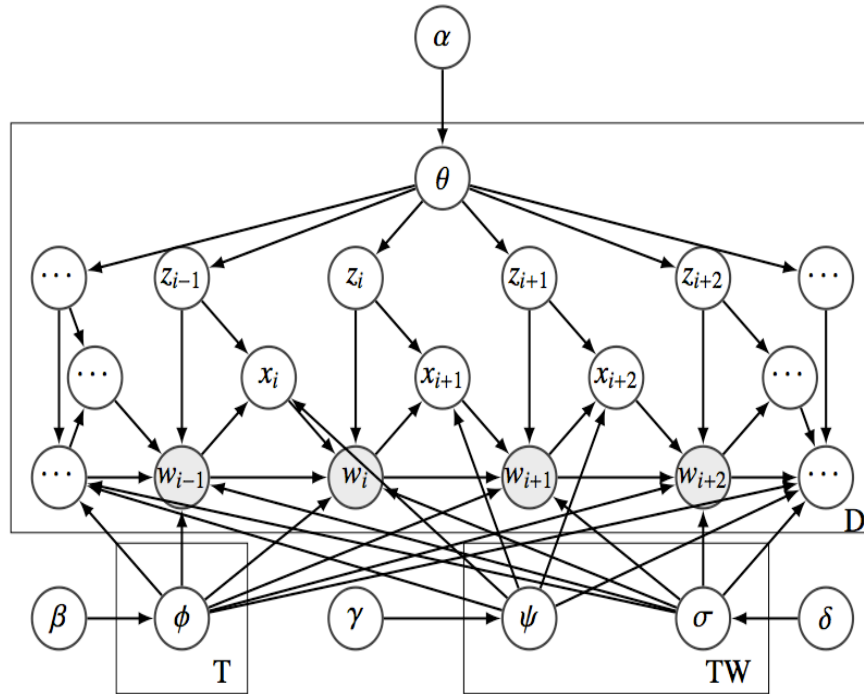


Figure 5: Graphical model of the topical n-gram Model

Figure 5 is a graphical model representation of the topical n-gram model (TNG), where D is the document level, T is the topic level, and W is the token level. Compared to the LDA model in Figure 4, the bigram status Bernoulli distribution ψ and the bigram distribution σ are new in Figure 5. Hence, more uncertainties are added to the joint probability $p(w, z, x | \alpha, \beta, \gamma, \delta)$. Gibbs sampling [25] is used to conduct approximate posterior inference in the topical n-gram model.

In the topical n-gram model, the last term of the n-gram is the word considered when generating the topics. That is, even though the topical n-gram model approach enriches the generated topics by longer sequences of words, the topic generation process is still based on individual words with the word context providing evidence to form a longer n-gram. As mentioned in Chapter 1, constituent terms cannot capture the rich meaning of the whole phrase. Besides, based on this approach, there is no evidence to remove high frequency n-grams that may not be important (eg., “tend to show”).

2.4 C-value Method

Extractive text summarization is an approach where short summaries of a collection of documents are generated by selecting a few sentences or phrases from those documents that represent the gist of the collection in some way. The C-value [12] method is an extractive text summarization method that extracts key phrases that capture a summary of a collection of documents. It uses both linguistic information [31] [32] and statistical information [33] [34] to identify the key phrases.

First the following three noun phrase regular expression filters are used to extract candidate phrases:

1. *Noun*Noun*
2. *(Adj | Noun)+ Noun*
3. *((Adj | Noun)+ | ((Adj | Noun)* (NounPrep?)(Adj | Noun)*)Noun*

Here *Adj* stands for adjective and *NounPrep* stands for a noun followed by a preposition. + means zero or more, * means one or more and | means logical “or”.

Next for each candidate phrase, the C-value is computed based on its frequency and the frequencies of longer phrases that contain it in the given set of documents. The C-value formula can be written as

$$C(p) = \begin{cases} \log_2(\text{len}(p)) \cdot f(p) & \text{if } p \text{ is not nested} \\ \log_2(\text{len}(p)) \cdot \left(f(p) - \frac{1}{|T_p|} \sum_{q \in T_p} f(q) \right) & \text{if } p \text{ is nested} \end{cases}$$

where $C(p)$ is the C-value of phrase p , $\text{len}(p)$ is the number of words in phrase p , and T_p is the set of the longer noun phrases that contain phrase p , and $f(p)$ is the frequency of p in all the documents of the corpus. If p is not nested, it implies that it does not appear in longer phrases. When it is nested, we discount its C-value based on the number of its occurrences in its longer phrases (the $\sum_{q \in T_p} f(q)$ part) and dampen this discount based on the number of unique longer phrases that contain it (the $\frac{1}{|T_p|}$ part). With

this measure, the larger the C-value, the more important is the phrase relative other phrases with lower C-value.

For example, Table 2.2 lists some phrase and their corresponding frequencies. “Real time” is a phrase nested in 5 unique longer phrases. Based on frequency alone, “Real time” seems more important than “Real time clock”. However, if one document contains these 5 phrases, “Real time clock” will be more important than the phrase “Real time”. If we calculate their C-value, we get

$$C(\textit{Real time}) = \log_2(2) \cdot (10 - \frac{1}{5} \cdot 10) = 4$$

$$C(\textit{Real time clock}) = \log_2(3) \cdot 6 \approx 4.17$$

Based on their C-values, “Real time clock” will rank higher than “Real time”.

Table 2: Example for C-value method

Phrase	Frequency
“Real time clock”	6
“Real time system”	1
“Real time output”	1
“Real time expert system”	1
“Real time imagegeneration”	1

2.5 Topic Browser

Given that topic models have great potential to unveil the hidden semantic structure under a large collection of documents as well as each single document, visualizing the result of topic models is an interesting and promising research topic. As Blei said in 2011,

“...Topic modeling algorithms show much promise for uncovering meaningful thematic structure in large collections of documents. But making this structure useful requires careful attention to information visualization and the corresponding user interfaces...” [35]

In 2012, Chaney and Blei developed the Topic Browser [14] to visualize the results of the topic models. In their Topic Browser, they summarize the corpus by displaying all the generated topics. Each topic is represented by showing the first three most prevalent words (or phrases in the case of our model) in the topic.

Here is an example in Figure 6. Displaying these topics can help users narrow down their interest to one or two particular topics. After users find some topic interesting, this browser allows them to navigate to the corresponding topic page, which could reveal the relationships between topic and documents and between topics and other topics. Figure 7 shows a topic page. The title of this page is the first three phrases in the selected topic. The left-hand column is the distribution of all phrases in the topic with the phrases ordered from highest to lowest probability. The higher the phrase is, the higher the probability the phrase has in this topic. The middle column is a list of document titles, ordered

by the probability that the paper is about this topic. The higher the document is, the more the document is related to (or about) this topic. The right column is the related topics ordered by their similarity to the selected topic. After reviewing the listed document title under this topic, the users may find some document interesting. Clicking on the document title displays the document page. This page can reveal the relationships between the document and topics and documents that are similar to the selected document. Figure 8 shows a document page. The left column shows the topics related to this document. These topics are also displayed in a pie chart, representing their proportions in that document respectively. The right column shows documents that are similar to this one in decreasing order of similarity.

Figure 6, Figure 7 and Figure 8 are all build based on 12,751 documents fetched from PubMed using the query “*drugs abuse*” within 5 years.

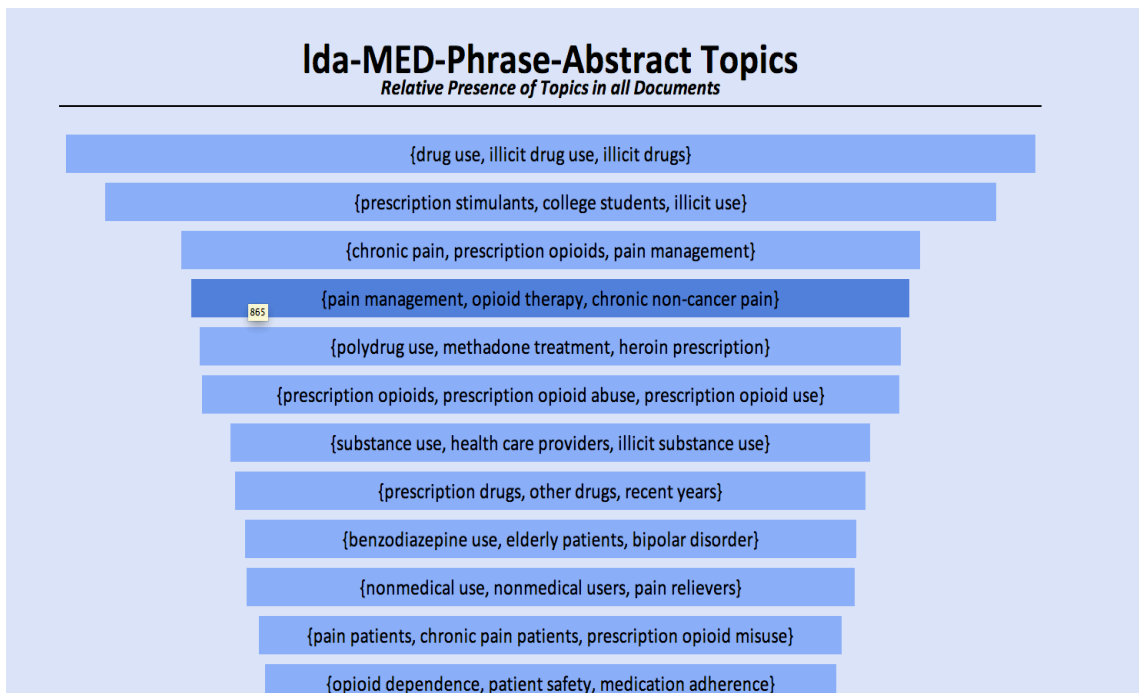


Figure 6: Topics summarization page of the Topic Browser

{pain management, opioid therapy, chronic non-cancer pain}		
words	related documents	related topics
pain management	Prescription Drug Abuse & Diversion: Role of the Pain Clinic.	{chronic pain, prescription opioids, pain management}
opioid therapy	Prescription for danger. Prescription drug abuse is rampant among today's adolescents; here's what you need to know about this epidemic.	{opioid analgesics, past decade, prescription opioid analgesics}
chronic non-cancer pain	In vitro characterization of ephedrine-related stereoisomers at biogenic amine transporters and the receptorome reveals selective actions as norepinephrine transporter substrates.	{prescription drug abuse, substance use disorders, substance use disorder}
opioid abuse	Characterizing the subjective, psychomotor, and physiological effects of oral oxycodone in non-drug-abusing volunteers.	{substance abuse, health problems, mental health}
opioid use	Prescription drug use and abuse. Risk factors, red flags, and prevention strategies.	{data collection, other medications, fatal poisonings}
pain conditions	The addicted physician. A rational response to an irrational disease.	{opioid dependence, opioid medications, primary care}
prescription opioid abuse	Patient characteristics associated with opioid versus	{prescription drug, response rate, young people}
chronic opioid therapy		
opioid misuse		
side effects		
opioid treatment		

Figure 7: Topic page of the Topic browser

Prescription drug use and abuse. Risk factors, red flags, and prevention strategies.


	<p>Prescription drug use and abuse. Risk factors, red flags, and prevention strategies.</p> <p>Isaacson JH, Hopper JA, Alford DP, Parran T.</p> <p>Source</p> <p>Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, OH 44195, USA. isaacsj@ccf.org</p> <p>Abstract</p> <p>When a patient is in physical or emotional pain, prescribing controlled substances often appears to be the simplest and most efficient way to relieve suffering and distress. However, in a minority of cases, this approach leads to prescription drug abuse and patient harm. In this article, the authors review the epidemiologic factors of prescription drug abuse, legal policies designed to safeguard against it, risk factors and red flags, and practical ways to minimize the chances of misuse.</p> <p>Full article ▶</p>	<p>related documents</p> <p>Characterizing the subjective, psychomotor, and physiological effects of oral oxycodone in non-drug-abusing volunteers.</p> <p>The addicted physician. A rational response to an irrational disease.</p> <p>Patient characteristics associated with opioid versus nonsteroidal anti-inflammatory drug management of chronic low back pain.</p> <p>[Most frequent drug-related events detected by pharmacists during prescription analysis in a university hospital].</p> <p>In vitro characterization of ephedrine-related stereoisomers at biogenic amine transporters and the receptorome reveals selective actions as norepinephrine transporter</p>
<p>related topics</p> <p>{pain management, opioid therapy, chronic non-cancer pain}</p> <p>{chronic pain, prescription opioids, pain management}</p> <p>{prescription drug, response rate, young people}</p>		

Figure 8: Document page of the Topic Browser

Chapter 3 Methods

In this Chapter, we describe the construction of our Phrase LDA model, the model evaluation and the interface evaluation that we conducted.

3.1 Phrase LDA method

We use the traditional LDA method by reducing the contents of documents to noun phrases for which the C-value computed over the set of documents to be modeled is greater than 2. This threshold for the C-value was determined based on our experimental analysis. Most of the phrases with C-value less than 2 appear only 1 time over the whole corpus, which are meaningless for topic model to generate topics and extract relationships between documents. In addition, removing these phrases can make the documents-phrases matrix more concrete and the computation of topic model more quickly. We also removed noun phrases longer than 10 words. According to our experimental analysis, once a phrase's length is longer than 10, there is a high probability that it's not a phrase. Note that phrases that occur multiple times in the same document are used as many times as they appear, that is duplicate are retained.

3.2 Evaluation of Models

To evaluate the phrase-based model we conducted an experiment with 11 participants using the word (phrase) intrusion test [13] and compared the Phrase LDA model with topical n-gram model. We obtained a corpus of 26,533 citations using PubMed query

public health[majr] AND united states[mh] AND "last 4 years"[dp]

to fetch the titles and abstracts from PubMed. This query fetches citations corresponding to articles that discuss public health as a major topic with US as a geographic location in the last four years. We chose this particular query as our participants are from the college of public health. We applied the time period constraint to limit the number of abstracts to a reasonable size. We treated each title and its corresponding abstract as a document. We first computed the C-value of the phrases from the corpus and retained only those phrases for each citation with the C-value larger than two. Based on this threshold, we chose 51,627 unique phrases out of the total 365,156 phrases. Next the text for each citation is replaced with the C-value > 2 noun phrases (including duplicates) that appear in the citation (abstract and title) text.

Figure 9 is a comparison between the unigram (or word) frequency distribution and the phrase frequency distribution computed for the Phrase LDA method. For the regular LDA, we chose 25,789 unique words out of 67,775 words based on the frequency threshold of two. From Figure 9, we can see that most words or phrases are located in the frequency range $2 \leq f \leq 100$.

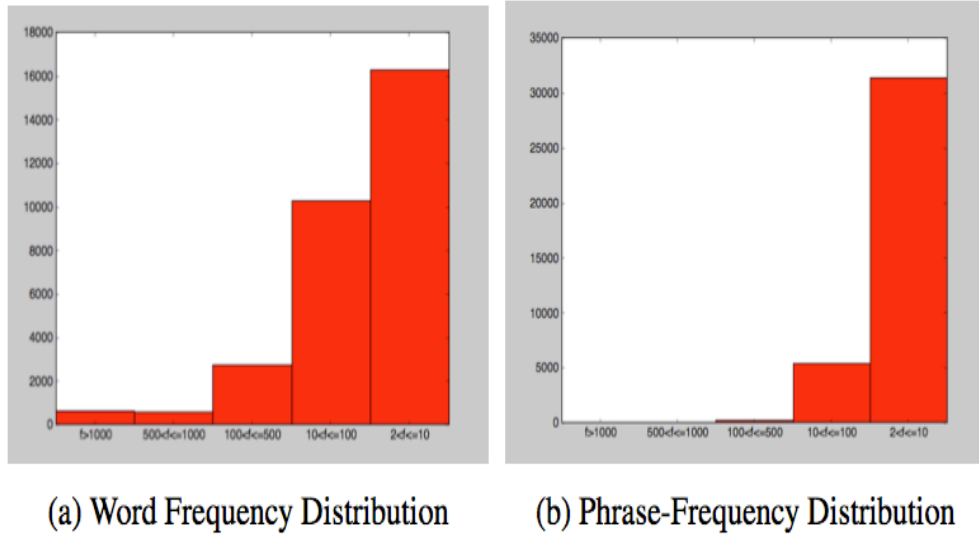


Figure 9: Comparison of Word and Phrases selected Distributions

We built topic models for our corpus using the general LDA, topical n-gram model, and our phrase LDA model. For the original LDA and Phrase LDA we used the implementation called LDA-C [36]. We set the number of topics at 50 for all models. The maximum number of iterations was 1000. We used the MALLET [37] toolbox for the implementation of the topical n-gram model. We set 50 topics and 1000 iterations for the 26,533 documents. A sample of topics generated by these three models is shown in Table 1. As can be seen, the n-gram models might not contain noun phrases and might just have frequent n-grams that are not necessarily meaningful. For example, “article includes” in the first topic in Table 3(b) is not a meaningful phrase.

Table 3: Three sample topics generated by LDA, topical n-gram model, and Phrase LDA.

Top ten words or phrase are listed for each topic.

Topic 1	Topic 2	Topic 3
quality	clinical	cost
medical	trials	costs
electronic	trial	per
data	studies	life
information	treatment	effectiveness
records	randomized	economic
health	study	model
record	research	million
care	results	treatment
patient	controlled	benefits

(a) LDA

Topic 1	Topic 2	Topic 3
health care	clinical modification	birth defects
public policy	diagnostic mammography	birth defect
health policy	distraction index	adverse pregnancy
public health policy	screening parameters	congenital heart defects
health research	risk tool	maternal deaths
based medicine	negative rate	multiple births
public health research	nosocomial cdi	assisted reproductive technology
article concludes	weight scaling	home births
medical decision	clinical examinations	control mothers
article argues	study subjects	maternal death

(b) Topical n-gram Model

Topic 1	Topic 2	Topic 3
newborn screening	health disparities	air pollution
clinical research	high rates	data sets
clinical studies	health care professionals	air pollutants
association study	social determinants	measurement error
trial results	health problem	epidemiological studies
health programs	health interventions	regression models
data elements	public health	ambient air pollution
health research	public health interventions	ambient concentrations
final rule	health care facilities	emission factors
screening programs	health inequities	aerodynamic diameter

(c) Phrase-Based LDA

3.3 Phrase/n-gram Intrusion Test

As topic models become more and more popular for the unsupervised analysis of large document collections [38], a number of advanced topic models have been developed to help people understand the hidden semantic structure of documents. However, given the the results of topic models are topics, document-topics distribution and topic-words distribution, there is no easy analytical way to determine whether topics generated by one model is better than another.

The probability of held-out documents given some training documents is often used as the secondary task to evaluate the topic models [39]. A better model will give a higher probability to the held-out documents. Unfortunately, extracting this probability is always intractable. Hence, several estimators have been developed. Some approaches include, importance sampling methods [40], harmonic mean method [41], annealed importance sampling [42], chib-style estimation [43], and “Left-to-right” evaluation algorithm [44].

However, topic models are developed to help humans understand large document collections. The evaluation of the model itself cannot guarantee that the topics generated by the models are better suited for this task. In 2009, Chang et al. [13] introduced an important intrinsic evaluation method called *word intrusion* for topic models that is independent of the application context. It involves using human subjects to evaluate the intrinsic coherence of the topics generated. We extended this to “Phrase/n-gram Intrusion” test to compare the quality of the topics generated by our model and the topical

n-gram model. We chose to leave out comparisons with the original word based LDA model because of the semantic diversity present in single words that sometimes provides an unfair advantage to word models in an intrusion identification method.

We randomly chose 25 topics out of 50 topics generated by the topical n-gram model and our phrase-based LDA model. For each selected topic, we then chose the top three phrases and randomly select one phrase out of the bottom five phrases as the *intruder phrase*. We randomly ordered these four phrases and presented it to the participant as a multiple choice question where the objective is to identify the intruder phrase. If the topics are semantically cohesive and meaningful, participants should be able to easily identify the *intruder phrase*. If the topics are incohesive, users might find it difficult to identify the intruder phrase and may resort to guessing. We built an anonymous questionnaire (https://uky.qualtrics.com/SE/?SID=SV_3qlfcTrBN6aNZt3) based on this phrase intrusion approach through the online survey software program Qualtrics (Qualtrics: <http://www.qualtrics.com/>). This questionnaire contains fifty questions and each question comes from one of the randomly selected 25 public health topics described earlier using the topical n-gram model and our phrase LDA model. Figure 10 is the example of this questionnaire. To make sure that each questionnaire is endowed with a minimal level of user concentration and reading comprehension, we added several simple questions (e.g., a question with choices {Father, Mother, Brother, Cancer}) to the questionnaire. If a user got any one of these simple questions wrong, we exclude this response from our analysis.

In [13], model precision is defined as

$$MP_k^m = \frac{1}{S} \sum_s 1(i_{k,s}^m = w_k^m)$$

where MP_k^m is the precision of model m for topic k , $i_{k,s}^m$ is the intruder phrase selected by user s for the topic k and model m , w_k^m is the actual intruder phrase selected by us for model m for topic k , and S is the total number of the subjects. The function $1(\langle condition \rangle)$ is a Boolean function that results in a 1 if condition evaluates to TRUE and returns a 0 if condition evaluates to FALSE. To compute the overall performance of a model, we calculate the average model precision as follows

$$AMP^m = \frac{1}{T} \sum_{k \in T} MP_k^m$$

where T is the total number of selected topics in model m .

3.4 Evaluation of Topic Browser

Information visualizations are difficult to evaluate, because they are primarily tools for supporting a creative process for developing insight and generating and then exploring hypotheses using open-ended discovery [16]. Thus a key measure of success of visualizations is whether they help biomedical researchers develop new questions and new hypotheses, not to simply answer pre-existing questions. Given topic models are developed to automatically summarize, organize and understand large collections of documents, the evaluation of this interface should focus on whether it will help biomedical researchers fulfill these goals.

To evaluate our tools, we will use the qualitative evaluation methodology developed by Saralya, North, and Duca for evaluating how well microarray visualization tools enabled biological insight. Three subjects used our Topic Model Visualization Browser based on their respective areas of research interest. Subjects were all experts in their own research areas. One subject interest's was "*prescription drug abuse*" with a total of 2649 records (titles and abstracts) related with this interest fetched from PubMed. Another subject supplied the PubMed query "*(((((((back [Title/Abstract]) OR trunk [Title/Abstract]) OR spine [Title/Abstract]) OR lumbar [Title/Abstract]) OR vertebral column [Title/Abstract])) AND (((biomechanic*[Title/Abstract]) OR mechanic*[Title/Abstract]) OR load*[Title/Abstract]) OR stability [Title/Abstract]))))*" resulting in 21,041 records fetched from PubMed based on this query. The last subject viewed documents from the PubMed query "*myositis AND ("skeletal muscle" OR macrophages OR inflammation OR regeneration) AND (Dermatomyositis OR "idiopathic inflammatory myopathy" OR polymyositis OR "inclusion body myositis" OR "cancer associated myositis")*" resulting in 1549 records.

All subjects were given 15 minutes of instruction and demonstration on how to use the tools along with a list of the kinds of questions that could be explored with the tools. This was designed to replicate the natural process whereby researchers learn to use new tools from other colleagues. Subjects were then instructed to list some questions they would typically ask about the data in the dataset, such as "Can you get a brief idea of what these documents are talking about?" and "Do these topics make sense to you?" and so on. After this, they were instructed to continue to use the tools to explore the dataset

until they felt that they would not gain further insight.

During the sessions, the subjects' comments were noted on pen and paper by the experimenter. We then analyzed the notes to extract the following dependent variables: users' motivation, total time spent with the tools, answers of list of initial questions, list of further insights, visualization techniques used, usability issues, and participant demographics [16].

What is your age?

- <18
- 18-21
- 22-25
- 26-29
- >29

Please select the term or phrase that does **not** belong with the others. For example, here is a list of ordered words, ('dog', 'pig', 'apple', 'cow'), as you can see, the 'apple' does not belong with the others, so it will be picked out. If there is no obvious selection, please select the one you think is most different. Please keep in mind that these choices are generated by the topic "Healthcare and US"

- chronic disease outcomes
- clinical studies
- clinical research
- newborn screening

Please select the term or phrase that does **not** belong with the others.

- air pollutants
- accurate determination
- data sets
- air pollution

Please select the term or phrase that does **not** belong with the others.

- public health
- health education
- current research
- medical education

Please select the term or phrase that does **not** belong with the others.

- emergency department
- trauma center
- reliability coefficients
- trauma patients

Please select the term or phrase that does **not** belong with the others.

- staff members
- blood samples
- cigarette smoking
- vertical reference

Figure 10: Example of the questionnaire. Each question has 4 choices and only one of them is the intruder.

Chapter 4 Results

4.1 Results of Models' evaluation

11 subjects completed the intruder phrase recognition questionnaire online. All these users are graduate students from different departments, for example, department of Computer Science, department of Electrical and Computer Engineering, and department of Pharmacy. But all of them are working on public health topics at the University of Kentucky. Five of them are in the age group 22-25, four in the age group 26-29 and the remaining three are at least 30 years old. Time spent on this questionnaire ranged from 5 to 45 minutes. The average time was about 20 minutes.

The model precision for topical n-gram model for the 25 topics was 48% and for the phrase LDA was 55%. Hence our Phrase LDA achieved a 7% improvement over the topical n-gram model based on this intrinsic evaluation. Furthermore, 8 of the 11 subjects indicated that the topics generated by our Phrase LDA model were easier to understand than those generated by the topical n-gram model.

The results show that our adaptation of the original word-based LDA to key phrases-based LDA resulted in better topic cohesion and improved comprehension when compared to the topical n-gram model. Hence, the C-value method improves overall comprehension while maintaining cohesion.

4.2 Results of Topic Browser's evaluation

3 subjects completed the evaluation of the topic browsers we built for them based

on their interests. Table 3 presents the details about the three evaluations of the topic browser.

Figure 11 is the 60 topics summarization page for the 2649 documents fetched from PubMed with the PubMed query “*prescription drug abuse*” for subject 1. From the list of the topics, the subject quickly navigated to the topic that he was interested in and found the interesting documents. One of these documents was the target document he found useful before using the topic browser, which confirms that Topic Browser can help for users locate the documents related to their research. After quickly reviewing other documents under these topics, the subject confirmed his suspicion that little is published in the area of interest (effect of drug screening programs on drug abuse).

Figure 12 is the 70 topics summarization page for subject 2 with the theme “*Back pain and biomechanic*”. The subject concluded that he could get a brief idea of what these documents were talking about based on these listed topics. Besides, he could also quickly determine the field of research behind each topic. Here are several examples from this subject, “low back pain, risk factors, work load” suggests studies conducted in the area of occupational biomechanics, ergonomics, and epidemiology that have an emphasis on prevention; “Back pain, low back pain, chronic back pain, pain patient, mechanical low back pain” suggests studies conducted by people in the area of health science like physical therapy, with an emphasis on rehabilitation; “muscle activity, muscle forces, lumbar spine, shear forces, trunk muscles” suggests an engineering approach to trunk biomechanics. This subject found 59 of 70 topics meaningful and also was able to identify synonymous across different research fields.

Figure 13 is the 40 topics summarization page for subject 3 with the theme “*myositis*”. Given that this subject had already done considerable research in this area, this subject was quite familiar the documents. After reviewing this Topic Browser, this subject concluded that this topic browser captured most aspects of this research area “*myositis*”.

All these subjects reported the following observations after using the Topic Browser based on their research interests.

- 1). Advantages: The subjects found the topic browser interesting to explore. They also noted that this tool helped them save a lot time reviewing the documents that they were interested in. Besides, this tool is helpful for these subjects to avoid the misunderstandings when ideas are being discussed by researchers in other related fields.
- 2). Improvement suggested: All the subjects felt that this tool needs a way to show the documents based on a combination of topics that they are interested in. What’s more, they were more likely to start with phrases instead of reviewing the topics one by one. Hence this tool needs a better way to help users to navigate from the phrases to the interesting topics.

Table 4: Three evaluations of the Topic Browser

	Subject1	Subject2	Subject3
PubMed Query Theme (Please see Chapter 3.3 for completed query)	<i>“prescription drug abuse”</i>	<i>“Back Pain and Biomechanic”</i>	<i>“myositis”</i>
Number of Documents	2,649	21,041	1,549
Motivation	To locate the interesting articles about the impact of drug screening programs on drug abuse.	Doing research on Back Pain in Biomechanics domain	Doing research on myositis
Time	30 minutes	45 minutes	30 minutes
Conclusion	<p>1, Helped the subject quickly review all the interesting documents</p> <p>2, Found several interesting articles</p> <p>3, Confirmed suspicion that little is published in the area of interest (effect of drug screening programs on drug abuse)</p>	<p>1, Helped the subject quickly review all the interesting documents</p> <p>2, Quickly got ideas about the research that is being done by people in other related fields.</p>	Given the subject has already done a lot research on the myositis, this topic browser confirmed this subject’s understanding of this topic

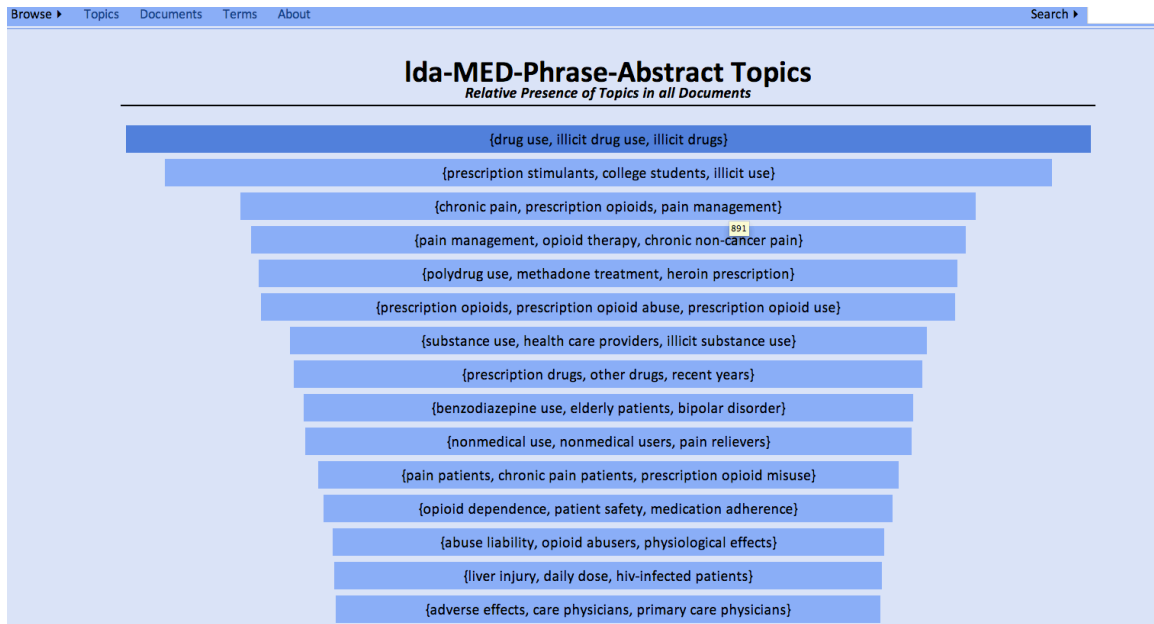


Figure 11: Topics summarization page for subject 1

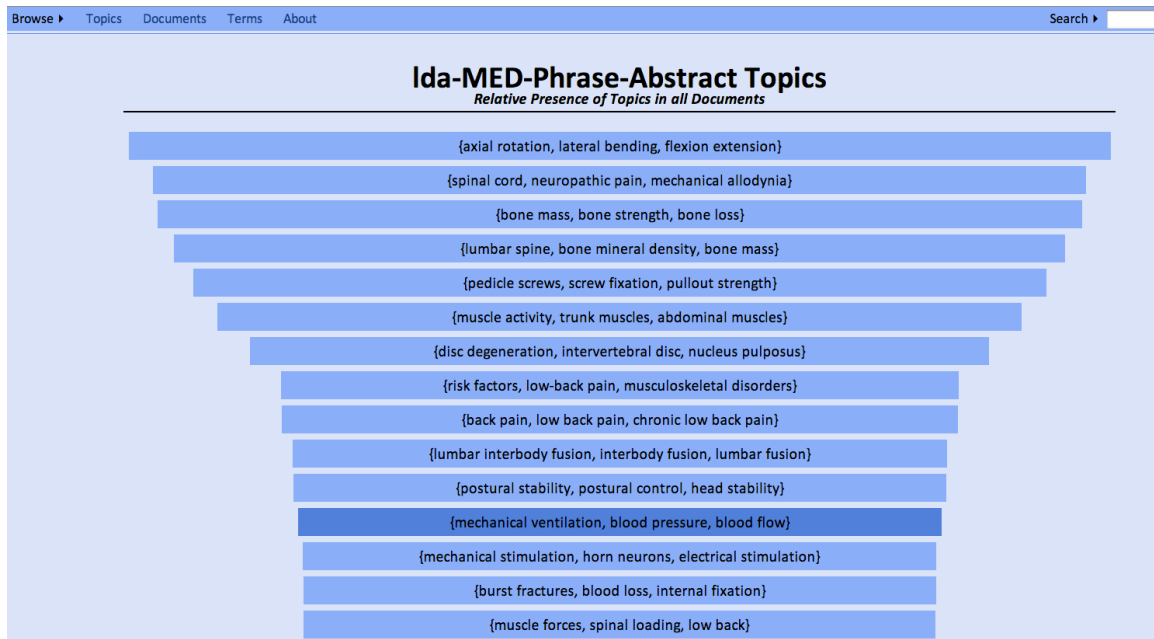


Figure 12: Topics summarization page for subject 2

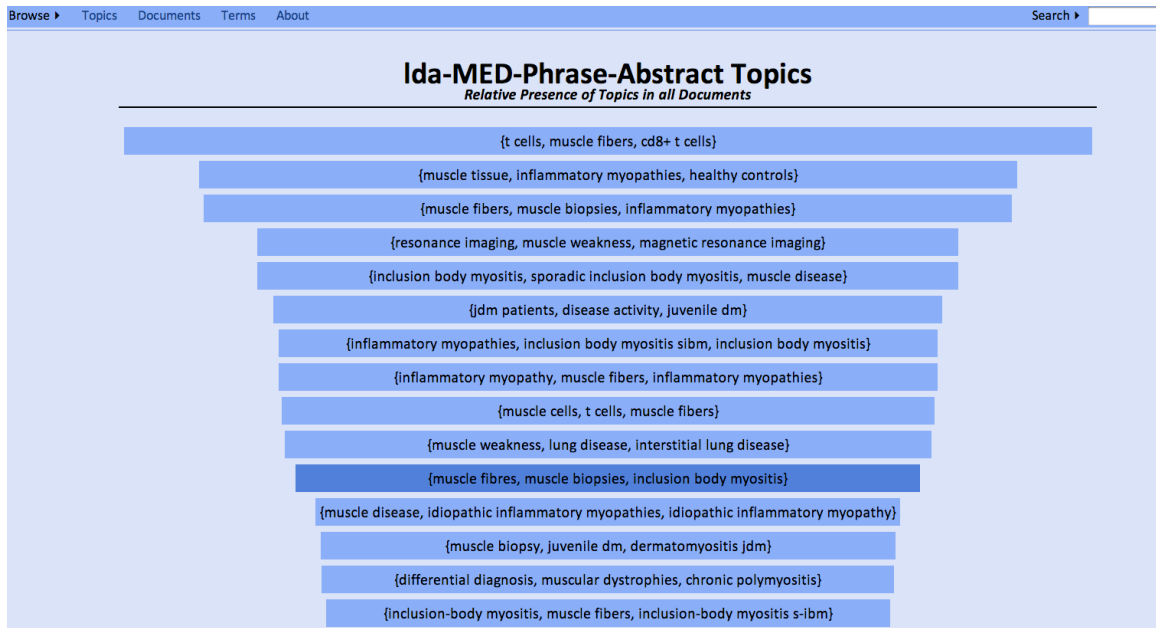


Figure 13: Topics summarization page for subject 3

Chapter 5 Conclusion and Future Work

5.1 Conclusion

In this paper we presented we presented this biomedical semantic information processing system build based on LDA model. In this system, we used a phrase based LDA topic modeling approach for biomedical documents and conducted intrinsic model evaluation through user study, which resulted in 7% improvement in model precision. We also conducted an empirical user study to evaluate the topic browser [14] interface we used in our system. The result shows that this system helps the users save time to search, review and understand the documents fetched from PubMed by showing the semantic structure under these documents in a Topic Browser.

5.2 Future Work

We conclude with three future research directions:

- 1). Topic models built on increasing levels of abstraction (words, key phrases, named entities, relations) might provide better ways of surfacing important and possibly new undiscovered themes when applied to sets of research articles. We plan to explore the potential of named entity and relation based topic models in our future work.
- 2). In this paper we only conducted intrinsic evaluation based on a user study. For future work, we plan to conduct extrinsic application based evaluation by using model parameters as feature weights in machine learning algorithms for text classification based on our Phrase LDA approach.
- 3). We are planning to build a more interactive Web based topic browser [14] for open-ended knowledge discovery, which could give users multiple ways of visualizing the large collection documents that they are interested in. Besides, the users can also have the rights to modify the topic browser we built for them. For ex-

ample, they can also delete unrelated documents, phrases, and topics; the user can choose to show the related documents based the phrases and topics combination that they are interested in.

Topic models have a great potential for analyzing the content of large text corpora. However, the deployment of topic models in the real world has been limited. Our targets in the future are to find ways to apply the topic models to help people better understand the digital data world.

Bibliography

- [1] Jinha, Arif E. "Article 50 million: an estimate of the number of scholarly articles in existence." Learned Publishing (2010): pp258-263.
- [2] Alan G Fraser, Frank D Dunstan. "On the impossibility of being expert." BMJ (2010): 1314-1315.
- [3] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine (1996): 37-54.
- [4] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of Machine Learning Research 3 (2003): 993-1022.
- [5] Atkins D, Rubin T, Steyvers M, Doeden M, Baucom B, Christensen A. "Topic Models: A Novel Method for Modeling Couple and Family Text Data ." Journal of family psychology: JFP: journal of the Division of Family Psychology of the American Psychological Association (Division 43) (2012): 816-827.
- [6] Lehman Lw, Saeed M, Long W, Lee J, Mark R. "Risk Stratification of ICU Patients Using Topic Models Inferred from Unstructured Progress Notes." AMIA Annual Symposium Proceedings. 2012. 505-511.
- [7] Blei, D., J. Lafferty. Text Mining: Theory and Applications, chap. Topic Models. Taylor and Francis, 2009.
- [8] Mimno, D., A. McCallum. "Organizing the OCA: learning faceted subjects from a library of digital books." JCDL (2007).
- [9] HM, Wallach. "Topic modeling: beyond bag-of-words." Proceedings of the 23rd international conference on Machine learning. 2006. 977-984.
- [10] MacKay, D. J. C., & Peto, L. C. B. "A hierarchical Dirichlet language model." Natural Language Engineering (1995): 289-307.
- [11] WangX, McCallumA,WeiX. "TopicalN-Grams:PhraseandTopicDiscovery,withanApplicationtoInformation Retrieval." Proceedings of the 2007 Seventh IEEE International Conference on Data Mining. 2007. 697-702.
- [12]Frantzi KT, Ananiadou S, Tsujii Ji. "The C-value/NC-value Method of Automatic Recognition for Multi-Word Terms ." Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries . 1998. 585-604.

- [13] Chang J, Boyd-Graber J, Wang C, Gerrish S, Blei DM. "Reading Tea Leaves: How Humans Interpret Topic Models ." NIPS. 2009.
- [14] Chaney AJ, Blei DM. "Visualizing topic models ." Association for the Advancement of Artificial Intelligence. 2012.
- [15] Blei, David M. "Introduction to Probabilistic Topic Models." Communications of the ACM (2011).
- [16] Shneiderman, B., Plaisant, C. "Strategies for Evaluating Information Visualization Tools: Multi-dimensional In-depth Long-term Case Studies." Proceedings of BELIV. 2006. 38-43.
- [17] Piatetsky-Shapiro, G and Matheus, C. "Knowledge discovery workbench for exploring business databases." International Journal of Intelligent Agents (1992): 675-686.
- [18] Papadimitriou, Christos, et al. "Latent Semantic Indexing: A probabilistic analysis." Proceedings of ACM PODS. 1998.
- [19] Richard O. Duda, Peter E. Hart, and David G. Stork. Pattern Classification. Wiley-Interscience. 2nd Edition. 2000.
- [20] Thijs Westerveld, Arjen P de Vries, Alex van Ballegooij, Franciska de Jong and Djoerd Hiemstra. "A Probabilistic Multimedia Retrieval Model and Its Evaluation." EURASIP Journal on Advances in Signal Processing (2003).
- [21] Jordan, Michael I. "Graphical models." Statistical Science: Special Issue on Bayesian Statistics (2003): 140-155.
- [22] Jordan, M., Ghahramani, Z., Jaakkola, T., & Saul, L. "Introduction to variational methods for graphical models." Machine Learning (1999): 183-233.
- [23] Blei, David M., and Michael I. Jordan. "Variational methods for the Dirichlet process." Proceedings of the twenty-first international conference on Machine learning. 2004.
- [24] Minka T. P., Lafferty J. D. "Expectation propagation for generative aspect model." Proc. 18th Conf. in Uncertainty in Artificial Intelligence. 2002.
- [25] Griffiths, T.L. , Steyvers. M. "Finding scientific topics." National Academy of Sciences of USA (2004): 5228-5235.
- [26] M. Hoffman, P. Cook, and D. Blei. "Bayesian spectral matching: Turning Young

- MC into MC Hammer via MCMC sampling." International Computer Music Conference. 2009.
- [27] A. Asuncion, M. Welling, P. Smyth, Y.W. Teh. "On Smoothing and Inference for Topic Models." Uncertainty in Artificial Intelligence (UAI). 2009.
- [28] Teh, Yee Whye. "A hierarchical Bayesian language model based on Pitman-Yor processes." Proceedings of the Association for Computational Linguistics. 2006.
- [29] Hoffman, Matthew, Blei, David M., and Bach, Francis. "Online learning for latent Dirichlet allocation." NIPS (2010).
- [30] Steyvers, M., and Griffiths, T. "Probabilistic topic models." Landauer, T.; McNamara, D.; Dennis, S.; and Kintsch, W., eds., Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum. (2006).
- [31] Ananiadou, Sophia. "A methodology for automatic term recognition." In Proceedings of the 15th Conference on Computational Linguistics. 1994. 1034-1038.
- [32] Bourigault, Didier. "Surface grammatical analysis for the extraction of terminological noun phrases." Proceedings of the 14th International Conference on Computational Linguistics. 1992. 977-981.
- [33] Ido Dagan, Church Ken. "Termight: Identifying and translating technical terminology." Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics. 1995. 34-40.
- [34] Chantal Enguehard, Laurent Pantera. "Automatic natural acquisition of a terminology." Journal of Quantitative Linguistics (1994): 27-32.
- [35] Blei, David M. "Introduction to Probabilistic Topic Models." Princeton University (2011).
- [36] Blei, D. Latent Dirichlet allocation. 2003. <<http://www.cs.princeton.edu/~blei/lda-c/index.html>>.
- [37] McCallum, AK. MALLET: A Machine Learning for Language Toolkit. 2002. <[Http://mallet.cs.umass.edu](http://mallet.cs.umass.edu). >.
- [38] Blei, D., J. Lafferty. Text Mining: Theory and Applications, chap. Topic Models. Taylor and Francis, 2009.
- [39] Wallach, Hanna M., et al. "Evaluation methods for topic models." Proceedings of the 26th International Conference on Machine Learning (ICML) (2009).

- [40] Li, W. McCallum, A. "Pachinko allocation: DAG-structured mixture models of topic correlations." Int'l. Conf. on Machine Learning. 2006. 577-584.
- [41] Newton, M. A. Raftery, A. E. "Approximate Bayesian inference with the weighted likelihood bootstrap." Royal Stat. Soc. B (1994): 3-48.
- [42] Neal, R. M. "Annealed importance sampling." Statistics and Computing (2001): 125-139.
- [43] Murray, I. Salakhutdinov, R. "Evaluating probabilities under high-dimensional latent variable models." Neural Information Processing Systems. 2009. 1137-1144.
- [44] Wallach, H. M. "Structured topic models for language." PhD Thesis, University of Cambridge. (2008).

Vita

Zhiguo Yu was born in Xuzhou City, Jiangsu Province, P. R. China.

Education

September, 2007 --- July, 2011

Bachelor of Information Engineering

Beijing Institute of Technology, P. R. China