An integrative computational approach to identify disease-specific networks from PubMed literature information

Yuji Zhang1, Dingchen Li1, Cui Tao2, Feichen Shen3, Hongfang Liu1
1Division of Biomedical Statistics and Informatics, Department of Health Sciences Research
Mayo Clinic, Rochester, MN
2School of Biomedical Informatics, University of Texas Health Science Center at Houston
Houston, TX
3School of Computing and Engineering, University of Missouri-Kansas City
Kansas City, MO
Email: {Zhang.Yuji, Li.Dingchen, Liu.Hongfang}@mayo.edu, Cui.Tao@uth.tmc.edu, fsm89@mail.umkc.edu

Abstract—A huge amount of association relationships among biological entities (e.g., diseases, drugs, and genes) are scattered in biomedical literature. How to extract and analyze such heterogeneous data still remains a challenging task for most researchers in the biomedical field. Natural language processing (NLP) has the potential in extracting associations among biological entities from literature. However, association information extracted through NLP can be large, noisy, and redundant which poses significant challenges to biomedical researchers to use such information. To address this challenge, we propose a computational framework to facilitate the use of NLP results. We apply Latent Dirichlet Allocation (LDA) to discover topics based on associations. The networks extracted from each topic provide a disease-specific network for downstream bioinformatics analysis of associations for each topic. We illustrated the framework through the construction of disease-specific networks from Semantic MEDLINE, an NLP-generated association database, followed by the analysis of network properties, such as hub nodes and degree distribution. The results demonstrate that (1) LDA-based approach can group related diseases into the same disease topic; (2) the disease-specific association network follows the scale-free network property, in which hub nodes are enriched in related diseases, genes and drugs.

Keywords—Latent Dirichlet Allocation; Semantic MEDLINE; Network Analysis; Disease-specific network

I. INTRODUCTION

Biomedical literature is a rich resource for biomedical scientists to explore associations among drugs, diseases, and genes. Natural language processing (NLP) has been applied to extract information from biomedical literature. However, it is computationally challenging to perform queries directly from these resources such as Semantic MEDLINE [1]. Furthermore, associations among different biomedical entities are very complex yet sparse, which is impossible for biological researchers to understand millions of associations associated with their studies. Informatics approaches are needed to fill gaps between information need of translational researchers and existing information discovery services. Meanwhile, Latent Dirichlet Allocation (LDA) emerges as a promising approach to analyze a large amount of data since it has the ability to find hidden associations and reduce the data dimension. However, it is challenging to present LDA results to biomedical researchers in a biologically meaningful way.

To enable the use of LDA for knowledge discovery for translational researchers, we proposed a two-step computational framework to address this challenge and illustrated it through the construction of disease-specific networks from Semantic MEDLINE. Specifically, we applied LDA to assign diseases into different topics based on their associations with drugs and genes in Semantic MEDLINE. We then investigated the associations among these diseases and their associated drugs and genes where a disease-drug-gene network was reconstructed for each disease topic. The properties of each network were investigated by their overall network property.

The organization of the paper is as follows. Section II provides background and related work. Section III introduces the data resources and the proposed two-step computational framework. Section IV illustrates the results generated from each step in the proposed computational framework. Section V provides a thorough discussion of the results and concludes the paper.

II. BACKGROUND AND RELATED WORK

In biomedical informatics, LDA has applications in diverse research fields. For example, Arnold et al. [2] applied LDA approach to identify clinically significant topics that can be used for case-based retrieval patients’ notes. In 2006, Angues et al. applied unsupervised LDA to primary clinical dialogues for visualizing shared content in communication [3]. Wang et al. developed BioLDA [4] to find complex biological relationships in PubMed articles. Wu et al. [5] made use of LDA to rank gene-drug relationships in biomedical literatures based on Kullback-Leibler (KL) distance between topics derived from LDA. Bisgin et al. [6, 7] mined FDA drug labels using topic modeling. Fifty-two unique topics, each containing a set of terms, were identified and then the probabilistic topic associations were used to measure the similarity between drugs. Bian et al. [8, 9] demonstrated that LDA can also serve as a powerful tool for drug repositioning. LDA-base approaches can also be used to interpret MeSH terms of literature [10]. All these studies demonstrated that LDA has great potential in biomedical informatics. Here, we proposed
to apply LDA-based approach to identify disease topics that include close disease terms based on Semantic MEDLINE knowledgebase.

III. MATERIALS AND METHODS

In this section, we describe the data resources used in this work. We then introduce our proposed computational framework that involves two steps: (1) construct disease topics by applying LDA topic modeling on Semantic MEDLINE, and (2) investigate disease-specific association networks in each disease topic. Fig. 1 illustrates the steps of the proposed approach.

Fig. 1. Overview of the proposed computational framework

A. Extraction of association data from Semantic MEDLINE

Semantic MEDLINE currently contains more than 56 million associations extracted from MEDLINE citations. The current database contains eight tables, including concepts, concept semantic types, concept translations, predication, predication arguments, and sentences. Different tables need to be joined in order to obtain information for a particular association between two entities. The database contains an all-embracing joined table that provides information about associations (source concept, predicate, and object concept), and their source PubMed IDs (PMIDs). It is a big challenge for individual researchers to make use of such huge and heterogeneous data. We have optimized and reorganized the relevant data in Semantic MEDLINE in Resource Description Framework (RDF). Using the Unified Medical Language System (UMLS) semantic types and groups [12], we extracted unique associations among drugs, diseases, and genes, and represented them in six RDF graphs [12]. Detailed description of this approach was described in our previous work [13]. These six tables were used as preliminary association data resources including all unique associations from PubMed literature. Disease CUIs involved in these tables were considered as disease information and used in LDA topic modeling step.

B. Latent Dirichlet Allocation

Latent Dirichlet Allocation is originally a text analysis approach. In this study, we applied it to grouping associations among disease, drug and genes. In the text mining analysis, LDA represents a document as a mixture of fixed topics. Under the context of our data shown as in Fig. 2, LDA represents a collection of PubMed abstracts as a mixture of fixed topics. Each topic z has a weight \( \theta^z_p \) in one abstract p. Each topic has a distribution over a finite vocabulary of concepts including disease, drug or genes. Each concept c has a probability \( \phi_c^z \) in topic z. Placing symmetric Dirichlet priors on \( \theta \) and \( \phi \), with \( \theta \sim \text{Dirichlet}(\alpha) \) and \( \phi^z \sim \text{Dirichlet}(\beta) \), where \( \alpha \) and \( \beta \) are hyper-parameters to control the sparsity of distributions, the generative model is given by:

\[
\begin{align*}
\psi_{ci}^z \sim \text{Discrete}(\phi^z), & \quad i = 1, \ldots, C \\
\phi^z \sim \text{Dirichlet}(\beta), & \quad z = 1, \ldots, K \\
z_i|\theta^p \sim \text{Discrete}(\theta^p), & \quad i = 1, \ldots, C \\
\theta^p \sim \text{Dirichlet}(\alpha), & \quad p = 1, \ldots, P
\end{align*}
\]

where \( K \) is the total number of topics, \( C \) is the total number of concepts in the patient collection, and \( \psi_{ci} \) and \( z_i \) are the passage and the topic of the \( i \)th concept \( c_i \) respectively. Each concept in the vocabulary \( c_i = \{c_1, c_2, \ldots, c_c\} \) is assigned to each latent topic variable \( z_i \). Given a topic \( z_i = k \), the expected posterior probability \( \beta^i \) of topic mixings of a given patient \( p \) and the expected posterior probabilities \( \phi_{ci}^p \) of code group \( c_i \) are calculated as below.

\[
\begin{align*}
\phi^z_{ci} &= \frac{n_{ci,k} + \beta}{\sum_j n_{ci,j} + C\beta} \\
\hat{\beta}^p &= \frac{n_{pk} + \alpha}{\sum_k n_{pj,k} + P\alpha}
\end{align*}
\]

where \( n_{ci,k} \) is the count of \( c_i \) in topic \( k \), and \( n_{p,k} \) is the count of topic \( k \) in patient \( p \).

In this study, we used the LDA approach to obtain the parameter \( \phi \) for every disease group. Relevant topics were extracted by using the R package topicmodels, which is based on Blei et al [14].

IV. RESULTS

A. LDA results overview

In this study, we selected 69,053 PubMed literature abstracts from Semantic MEDLINE. In total, 1994 disease CUIs were involved in these abstracts. Based on their associations with other biological entities, we grouped these disease CUIs into 100 disease topics. Among these 100 topics shown in Fig. 2, their proportion is quite different. The first topic occupies 0.019 while the 100th only 0.0065 among the associations. This is part due to possible biases (e.g., diseases that have been studied more are prone to have higher proportions) and incomplete knowledge in existing literature abstracts. The top nine topics occupy about 15%.
For the top nine disease topics (Fig. 3), we demonstrated that there are distinct distributions among the disease CUIs in each disease topic. As shown in Fig. 3, each topic has its own dominant concepts without much overlapping with each other. Namely, each topic is represented by a few key concepts (the key concepts are determined by thresholds, which we set up as 0.04 based on the results). In general, three patterns among those dominant concepts can be found among the 9 topics. Firstly, one concept comprise overwhelming majority and a few others still hold significant portions. For instance, in disease topic 1, the four concepts occupy about 70% of the whole. Among them, C0027651 has shared 0.58 alone. More remarkable is topic 4 in which C0000768 possesses 0.72 while C0080178 0.06. The second pattern is that one or two concepts own almost the same ratio as the summation of all other dominant concepts. Topic 2 and topic 6 fall into this category. Topic 2 is dominated by 6 concepts with both C0016658 and C0040034 occupying about 0.28 and 0.11 respectively and the other four together about 0.25. The third pattern is that all the dominant concepts are evenly distributed. For example, in topic 3, the four main concepts, together only share about 0.2 in total. But all other 1990 insignificant concepts have far lower proportions as those four. Topic 3, 5, 7 and 9 all share this pattern.

B. Network construction and analysis of top topics

For each disease topics generated from Section A, we constructed disease-specific association networks. For instance, in disease topic 1, Neoplasms has about 58% proportion, a dominant concept in this topic. The other three disease CUIs, Candidiasis, Leukemia, Myelocytic, Acute, and Acute leukemia are all related to Neoplasms. Invasive Candidiasis is a common and serious complication of cancer and its therapy [16, 17]. In acute Leukemia, chronic systemic candidiasis has been recognized with increasing frequency [18]. It is not surprising that the Neoplasms-related disease topic are ranked No. 1 out of 100 topics since it has been a huge amount of research publications focusing on this disease group. This result indicates that the LDA-based approach is capable to capture such associations among thousands of literature abstracts. Based on disease CUIs in each disease topic, we further reconstructed disease-specific association network by extracting associations related to these disease CUIs. It is impossible for domain experts to read the whole association list. We applied network-based approach to investigate these association networks. The properties of each network were investigated by their overall network property. We selected Topic 3 (3316 associations) to investigate the overall network properties. Four hub nodes were identified: Huntington Disease, Pleural effusion disorder, Pregnancy, Ectopic, and Kaposi Sarcoma. These four hub nodes have many common associated nodes, indicating that they have some biological commonality although there is no direct supporting evidence in literature so far. This case study clearly demonstrates the superiority of network-based approach in inferring indirect associations among diseases in a disease topic generated using LDA.

V. CONCLUSIONS

In this paper, we proposed a novel two-step computational approach to facilitate knowledge discovery utilizing biomedical literature. The first step applies LDA to discover topics based on associations. The second step involves network construction and analysis of associations for each topic. The investigation of disease-specific association networks demonstrates the approach is able to reveal the association patterns and identify new knowledge. The results demonstrate that (1) LDA-based approach can group related diseases into same disease topic; (2) the disease-specific association network follow the scale-free network property, in which hub nodes are enriched in related diseases, genes and drugs.

Based on the results from this study, we expect to extend this work in several related research topics in the future. Some future extensions of this work include but are not limited to: (1) integration of additional supervised information (e.g., key words for PubMed abstracts) to make LDA generate more controllable and interpretable topics; (2) integration of more comprehensive association databases among disease, drug, and gene (e.g., HPRD and Drugbank) to construct more complete base association networks; (3) a framework to automatically extract such disease-specific association network so that such analysis can be extended to each disease topic; (4) additional network-based investigation of the relationships among disease, drug, and gene at other network level such as module subnetwork identification; (5) investigation on possible ways to improve the network by...
assigning weights or confident rates to different types of associations or associations from different sources.

ACKNOWLEDGMENT

This project was supported by the National Science Foundation under Grant #0937060, the Cancer Prevention & Research Institute of Texas (CPRIT R1307) to C.T. and the National Institute of Health grant 5R01LM009959-02 to H.L.

REFERENCES


