EXTRACTING DISCRIMINATING FEATURES FROM CONTEXTUAL DOMAIN KNOWLEDGE FOR ENHANCING TOPIC DISCOVERY OF ONLINE FORUMS

A Thesis in
Information Sciences and Technology

by

Chong Zhou

© 2016 Chong Zhou

Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

May 2016
The thesis of Chong Zhou was reviewed and approved by the following:

John Yen  
Professor of Information Sciences and Technology  
Thesis Adviser

Prasenjit Mitra  
Professor of Information Sciences and Technology  
Thesis Adviser

Matthew Ferrari  
Assistant Professor of Biology

Andrea H. Tapia  
Associate Professor of Information Sciences and Technology  
Graduate Program Director of Information Sciences and Technology

*Signatures on file in the Graduate School.
Abstract

Topic models are popular in discovering latent topics from document corpora. Traditional topic models such as Latent Dirichlet Allocation (LDA) are usually fully unsupervised, which are often found to generate ambiguous or incoherent topics based on human evaluations. Many extensions of topic models have been proposed to address this issue by incorporating prior knowledge into topic model to guide the topic discovery. Such prior knowledge can usually be defined as the discriminating words or cannot-links between two words indicating that these words should not be put into the same topic.

However, discovering the domain prior knowledge is not an easy job. Common practice usually requires domain experts to manually define the discriminating words, which require laborious efforts and may even fail to fit the data closely. In this thesis, we explore the possibility to automatically extract discriminating words from relevant informal domain knowledge. For example, in epidemics online course forum, the discussion is usually related to the syllabus of the course. The course syllabus can serve as a source of informal domain knowledge that contains a structural information of topics and reflects the high level knowledge of the forum. We can call this kind of contextual domain knowledge as domain contexts.

To extract the discriminating words from the domain contexts automatically, we first extend the information distance measures to discover discriminating words with respect to the domain contexts. Specifically, we introduce KL Distance and Loss of Mutual Information to determine the most discriminating words that should be separated.
Using these discriminating words, we developed an agglomerative algorithm to construct discriminating feature sets, which can be easily encoded as the cannot-link constraints to improve topic models. Then we incorporate the constraints into a state-of-the-art Multi-generalized Pólya Urn model to extract topics. Finally, we introduce a technique to discover the relationships between the discovered topics and the original domain contexts by mapping the top-k topics to each domain context elements. Our experiments on three large forum data sets show that even without any user input, it is possible to extract meaningful discriminating words automatically and discover coherent topics.
Table of Contents

List of Tables .................................................. viii

List of Figures .................................................. ix

Acknowledgments ................................................ x

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

Chapter 2. Background and Related Work ......................... 7
  2.1 Information Measures ........................................ 7
  2.2 Topic Modeling ............................................. 8
  2.3 Latent Dirichlet Allocation ................................. 8
  2.4 Topic Model: learning with knowledge ....................... 8

Chapter 3. Overview of the Methodology ......................... 11

Chapter 4. Information Distance Measures ....................... 12
  4.1 Information Theory Review .................................. 12
    4.1.1 Entropy .............................................. 12
    4.1.2 Mutual Information .................................. 13
  4.2 Extracting Discriminating Features with Information Distance Measures ............................................. 14
    4.2.1 KL Divergence and KL Distance ....................... 16
4.2.2 Loss of Mutual Information ........................................ 17

Chapter 5. Discriminating Features Sets ................................ 20

Chapter 6. Incorporating the Domain Constraints into Topic Models .... 23

  6.1 Dealing with Cannot-Links ......................................... 23
  6.2 Sampling Technique .................................................. 24
    6.2.1 Pólya Urn Model ............................................... 24
    6.2.2 Simple Pólya Urn Model (SPU) ............................... 24
    6.2.3 Generalized Pólya Urn Model (GPU) ........................ 25
    6.2.4 Multi-generalized Pólya Urn Model (M-GPU) ............. 26

Chapter 7. Extracting Relationships between Derived Topics and Domain Contexts 28

  7.1 Bipartite Graph Relationships .................................... 28
  7.2 Gibbs Sampling ...................................................... 29
  7.3 Future Extensions in Information Retrieval ........................ 30

Chapter 8. Data set ....................................................... 32

  8.1 Epidemics Course Forum ......................................... 32
  8.2 Machine Learning Course Forum ................................ 33
  8.3 Breast Cancer Forum .............................................. 33
  8.4 Data Preprocessing ................................................ 34

Chapter 9. Experiments .................................................. 35

  9.1 Evaluating Discriminating Feature Sets .......................... 35
  9.2 Topic Coherence ..................................................... 39
9.3 Relationships between Derived Topics and Domain Contexts . . . . . 43

Chapter 10. Conclusions .................................................. 46

10.1 Future Work .......................................................... 46

Appendix. DFS and topics of machine learning data set ................. 48

Bibliography .............................................................. 50
List of Tables

9.1 Keywords picked by domain experts from each weeks’ syllabus of epidemics course ............................................. 36
9.2 Discriminating Feature Sets learned with KLD and LMI in epidemics course syllabus ............................................. 37
9.3 Weighted Average Consistency and Diversity of the discriminating feature sets .................................................. 38
9.4 General topics and emerging topics in epidemics course forum .......................................................... 44
A.1 Discriminating Feature Sets learned with KLD and LMI in machine learning course syllabus .................................. 48
A.2 20 topics in machine learning course forum discovered by the M-GPU+LMI model ............................................... 49
List of Figures

3.1 Work flow of the proposed methodology .............................. 11

4.1 In the epidemics forum data set, the probability distribution of syllabus structure C induced by social, network, and vaccine .......................... 15

5.1 Example of the first 5 iterations of algorithm 1 on the epidemics course forum data set .................................................. 20

6.1 Simple Pólya Urn Model ..................................................... 25

9.1 Topic Coherence scores on all datasets using topic number $k = 15 \sim 50$. 41

9.2 Relationships between derived topics and domain contexts in the epidemics course forum dataset ........................................... 44
Acknowledgments

I would like to give my sincere thanks to my committee members Dr. John Yen, Dr. Prasenjit Mitra, and Dr. Mathew Ferrari for their time and effort to make this research possible.

We also wish to thank Brian Lambert, for helping to gather the epidemics course forum data in this work. We are also grateful for the four teaching assistants of the epidemics course forum in providing helpful insights. Furthermore, we will give thanks to Dr. Kenneth Portier for the useful discussions regarding the topics of the breast cancer forum operated by American Cancer Society (ACS), as well as Dr. Cornelia Caragea and Yafei Wang for the inspiring discussions. Finally, I will give thanks to my wife Li Jiang, for her suggestions and support to improve my work.
Chapter 1

Introduction

As Web 2.0 applications developed, we have witnessed a lot of popular online social medias (such as forums, news groups, blogs, etc.) emerge and thrive. Among these technologies, online forums are a very unique type of information exchange platform. Typically, online forums have a particular set of topics of interest where users can share information or ask questions in a thread structure. Users can also comment or reply to the thread to respond and communicate to the each other. In such an environment, analyzing the topics in the discussion is usually a very interesting and important task. For example, in a discussion forum of a Massive Open Online Course (MOOC), students can ask questions and help each other to better learn the course. Sometimes, it is even beneficial for the teachers to know the topics of those discussions so that they can address those questions in the following classes or reply to the thread.

Traditionally, researchers have utilized topic models such as Latent Dirichlet Allocation (LDA) [8] and Probabilistic Latent Semantic Analysis (pLSA) [25] to discover hidden topics in text documents. Many extensions [6, 34, 52] of LDA and pLSA, are introduced during the past decades and showed their values. However, these traditional topic models often tend to mingle several different concepts with close meanings into a single topic. This phenomenon gets worse when the scope of the data is a narrow topic
itself. As a matter of fact, most unsupervised topic models may not be able to produce coherent topics that are consistent with human judgments[11].

To help topic models to create more meaningful topics, there have been a lot of studies that include auxiliary domain prior knowledge into topic models, either by incorporating the authorship of the documents[48], label and tags [47], syntax[9], or document classes[7]. Similarly, some knowledge-based models allow users to define domain prior knowledge as constraints to regulate the topic generation process[30, 2, 29].

However, few of the above methods can leverage the contextual knowledge of online forums to improve topic qualities. Unlike other conversational texts, the topics in online forums are usually narrow and highly focused on a certain domain, which brings unique challenges to topic modeling. For example, in an epidemics course forum, it is often difficult to distinguish different topics in such a narrow domain where almost all topics are about infectious diseases. Fortunately, some domain contexts such as the course syllabus that correlate with the topics of online forum discussions are often available. These domain contexts are usually created by domain experts to reflect the structure of the related domain, which has high information value. Therefore, it is our goal to leverage valuable domain knowledge from the domain contexts to enhance topic models.

Consider the following real world example in an epidemic course forum. The syllabus of the course outlines eight weeks of the course materials. The first week of the course provides an overview of epidemics with key words including ”pathogen”, ”disease” and ”virus”. The following weeks focus on various aspects such as ”hosts”, ”vaccination” and ”disease control”. However, we observe that the topics generated by LDA
often fail to separate different concepts and form mixed topics such as \{disease, host, virus\}, \{vaccination, control\}, which does not help the lecturers and students to better understand the forum structure. It is expected that the generated topics are more fine grained and have specific meanings. Furthermore, only after we correctly discover the topics that correlate with the course content, we can then find emerging topics in the forum that are beyond the scope of domain contexts, which could help to facilitate discussions of these topics in the future.

The above example inspired us to construct some constraints in topic models and create a separation of certain key words. For example, we may construct a set of discriminating features \"\{disease, host, vaccine, control\}\" to specify that any two of these features should not appear in a same topic. In this way, the generated topics could be more focused on a sub-domain. Similar constraints has been proven to be effective in the literature \[2, 60, 15\].

Although incorporating constraints into topic models has shown its potentials, obtaining the proper domain prior knowledge is usually not an easy job. Common practices require domain experts to define the constraints manually \[2, 30, 3, 60\], but it takes laborious work from the domain experts and is hard to scale. It becomes even harder to acquire prior knowledge when dealing with incremental data and evolving topics.

In this thesis, we propose to extract domain constraints from domain contexts automatically using information distance measures, which consists of four steps:
First, to learn whether two words should be separated into different topics, we propose to use a consistent information distance measure to extract the discriminating features. Clearly, learning the distance without the contexts of a domain could be mistaken especially in a narrow domain. We explored the information measures Kullback-Leibler Distance and Loss of Mutual Information to learn the information distance between words, using the domain contexts as input. Intuitively, by conditioning the information distance on domain context, the proposed measures can quantify the differences of probability distribution of domain contexts induced by the word pairs, which can be used to discover the discriminating words that are the most representative on different concepts by maximizing the information distance. This is further explained in chapter 4.

Second, while simply considering the maximum information distance between word pairs can be useful, it will also give us a lot of redundant information. Imagine if we want to distinguish the concepts in a syllabus for 10 weeks of a course, and if we choose one representative word from each week, we will end up with $C(10, 2) = 45$ different combinations between these word pairs, which is difficult for human to interpret and adjust. To tackle this problem, we introduce discriminating feature sets such that each word pairs within each set are mutually discriminating. We introduce an agglomerative algorithm based on the idea of abstract features[31, 61, 50] to construct discriminating feature sets greedily. Such domain constraints can be easily incorporated to split discriminating words into different topics and enhance topic modeling.

Third, we leverage a state-of-the-art Multi-generalized Pólya Urn Model (M-GPU)[12] to incorporate the extracted domain constraints into topic models. M-GPU
model is similar to LDA but uses a simple logic during the parameter inference process. When a word is assigned to a topic, then the model will try to throw the words that shares a cannot-link with this word out of this topic and put them into the topics that are more suitable. With this mechanism, we can safely incorporate the domain constraints as cannot-links into topic models.

Finally, we will need to extract the relationships between the derived topics and domain contexts. For example, when the teachers of the Epidemic forum provided the syllabus and derived a more clearly separated topical structure, they may want to know that which topic is correspond to which week in the syllabus and whether there are some emerging topics that the students are interested in but not taught in class. This mapping relationships can be extracted with a simple Gibbs sampling technique. We propose to use a simple Gibbs Sampling technique to inference the distribution of topics in each domain context element. And we select the top five topics for each domain context element to create a link and derive the mapping relationships.

In summary, this paper makes two main contributions:

(1) It leverages domain contexts for online forums to enhance topic models. To our knowledge, this is the first time that documents with small size but high information value are being used to enhance topic modeling.

(2) It employs information measures to automatic discover discriminating features that needs to be separated into different topics, which is not well studied before. This allows us to further construct discriminating feature sets where any pairs of features are mutually discriminating.
Our experiments on epidemics course forum and breast cancer forum show that the propose method can effectively discover domain constraints as discriminating feature sets and can generate significantly more coherent topics.
Chapter 2

Background and Related Work

2.1 Information Measures

The use of information measures is popular in data mining. A common practice is to use mutual information\cite{17} to discover the most important features that contribute to different class labels\cite{21, 43} in feature selections. We can also minimize the information distance between two features to select the most similar features and group them together to form clusters\cite{4, 46} or as abstract features\cite{50, 61, 31} to improve classifications.

KL divergence\cite{32} or JS divergence\cite{19} has also been used in topic modeling in the past. We can leverage KL divergence for parameter inferencing \cite{6}. It is also very common to use the divergence measures to discover similar topics \cite{26, 1, 58, 62, 39} since a topic is a probability distribution over words.

However, although minimizing the divergence will reveal similar distributions, maximizing the information distance will result in the most discriminating features \cite{18}. In our research, we take this idea to find the discriminating features with maximum information distance, which has not been done in the topic modeling field before. We also take a step forward to greedily combine the discriminating features into discriminating feature sets to better enhance topic models.
2.2 Topic Modeling

2.3 Latent Dirichlet Allocation

Topic model is a set of statistical models in machine learning to discover the hidden "theme", or "topics" in a document corpus. Conceptually, a topic is a mixture of concepts or meaningful words. For example, the topic "genetics" may contains "human", "genes", "dna", and "genomes". The main idea behind topic models is usually a generative process to describe how a document is generated. For example, a document is viewed as a statistical distribution over the topic space and each topic is a statistical distribution over the vocabulary. The most popular topic model is proposed by Blei et al. in 2003, known as Latent Dirichlet Allocation (LDA) [8].

Latent Dirichlet Allocation assumes that a document is a collection of exchangeable words, also referred as the bag-of-words assumption. Each word in the document is generated by 1) randomly choose a topic from the document-topic distribution of the document; 2) randomly choose a word from the topic-word distribution of the topic selected from step 1.

Due to the size of real word corpus, the exact inference of the model parameters is usually intractable. Common approach to estimate the parameters includes Variational Inference [8] and Gibbs Sampling [22].

2.4 Topic Model: learning with knowledge

Although the simple LDA model is powerful enough to discover the hidden topics in large amount of text, it still often fails to correlate with human judgments[11]. A
common way to overcome its shortcomings is to incorporate more information to adapt to a specific task or domain.

A popular idea is to include some metadata of the documents such as authorship[48], user-defined tags[47], syntax[9], document categories[7], sentiments[35, 24] or entities[23].

More recently, knowledge-based topic models are proposed to incorporate general domain knowledge as constraints into topic models.

The idea of constraints is originally coming from clustering as "must-links" and "cannot-links" [53]. A must-link between two words restrict that these words must be put into the same topic while a cannot-link constraint puts the words into different topics. Extending this idea, [2] proposed a dirichlet forest method to incorporate must-links and cannot-links on words; [29] and [28] proposed to include users to put must-links between words iteratively while training the topic model. [57] took a similar approach in iterative training but the constraints are imposed on document level. [15] further extends must-links and cannot-links into set of words. However, all of the above studies does not provide a general method to construct must-links and cannot-links constraints automatically and require manual efforts.

Some attempts had been made to extract the constraints automatically. [14] presented the possibility to use synonyms and antonyms, which mainly focus on adjective and adverbs. Similarly, [59] leverages a probabilistic knowledge base to extract constraints. In comparison, our method will better suit the domain since we directly extract the constraints from the domain. [60], [55] and [56] proposed rule-based heuristic approach to form the constraints specifically used for product features. However, some of their assumptions (e.g. if two features appear in the same sentence but not connected
with "and" form a cannot-link) will be unreasonable in other domain while our approach can be easily migrated to other domain data sets. [16, 13, 12] employ the idea of transfer learning to learn the constraints from all other domains. However, in the example topic \{vaccine, disease, virus\} in epidemic course forum, it’s ideal to separate them into two topics \{vaccine\} and \{disease, virus\}, but they may even be assigned with a must-link based on the knowledge learnt from other domain. We take an approach to learn the knowledge directly from the contextual domain knowledge of the target domain, which is very important to extract domain specific constraints.

Inspired by the works above, we believe that it is possible to extract the domain constraints automatically. We believe that compared to the raw data that may contain a lot of useless or even misleading information, the domain contexts data generated by domain experts contains higher information values and is more reliable to extract constraints such as must-links and cannot links. In this thesis, we explore and introduce the techniques to extract domain constraints automatically and use them to guide the topic generation process.
Chapter 3

Overview of the Methodology

Figure 3.1: Work flow of the proposed methodology

Figure 3.1 shows the work flow of the proposed method in this paper. It has four steps: 1) with the domain contexts of the online forum such as syllabus or wiki pages, we extract the discriminating features using information distance measures. 2) We employ an agglomerative algorithm to construct abstract domain constraints as discriminating feature sets, where any two features in a set should not be put in a same topic. 3) Using the discriminating feature sets as constraints, we plug them into the state-of-the-art Multi-generalized Pólya Urn model[12] to generate the topics in the online forum discussions. 4) By mapping the topics in the forum with the domain contexts, we aim to discover potential emerging topics and form a better understanding of the topics in online forums.
Chapter 4

Information Distance Measures

In this chapter, we first review several concepts in information theory that are the building blocks of our work. Then we introduce the information distance measures to discover discriminating words.

4.1 Information Theory Review

Originated from the Claude E. Shannon’s work in signal processing, information theory has been widely used in many fields such as mathematics, electrical engineering, computer science, etc. [17]. It attempts to quantify the information of data based on probabilities. It is the foundation of many feature selection methods such as information gain and mutual information.

4.1.1 Entropy

The most basic concept in information theory is entropy ($H$) [49]. It measures the amount of uncertainty that a random variable $X$ contains. $H(X)$ is defined by

$$H(X) = - \sum_{x \in X} p(x) \log_b p(x)$$

where $b$ is the base of the logarithm. We will take $b = 2$ in this thesis and the information is measured in bits. Note that entropy is only dependent on the probability of $X$ and
does not consider its actual values. Intuitively, it measures how uncertain we know about 
X. If X is definite, then $H(X) = 0$, while $H(X)$ will reach its maximum value when X 
is randomly distributed.

4.1.2 Mutual Information

The mutual information $I$ [17] between two random variables $X$ and $Y$, is 
defined as follows.

$$I(X;Y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

Mutual information quantifies the amount of shared information between $X$ and $Y$. Note 
that we compare $p(x_i, y_j)$ with $p(x_i)p(y_j)$ in the equation. If $X$ and $Y$ are independent, 
this value will become 0, which indicates that we cannot know anything about $X$ given 
$Y$. It achieves its maximum value when $X = Y$. Moreover, mutual information is 
non-negative (i.e. $I(X;Y) \geq 0$) and symmetric (i.e. $I(X;Y) = I(Y;X)$).

Mutual information between a word and the domain contexts can be used to 
indicate how well a word can represent the domain context. Using the epidemics course 
forum as an example, the syllabus of the course is considered as the domain contexts $C$. 
Suppose there are $k$ weeks of material so that the syllabus is naturally divided into $k$ 
categories, which mean $C = \{c_1, c_2, ..., c_k\}$. We can calculate which word has the highest 
prediction power about $C$ using $\max \{I(C;w_i)\}$.
Although mutual information is useful in selecting the representative words based on domain contexts, it does not help us to discover domain constraints such as cannot-links between words $w_i$ and $w_j$. Directly calculate the mutual information between two words as $I(w_i; w_j)$ may very likely to fail. On one hand, our domain contexts are usually so short that many top values of this measure are identical as many word pairs seem to have the same probability distributions. On the other hand, $I(w_i; w_j)$ does not give considerations to the domain contexts $C$ and thus is not tuned to fit the domain.

4.2 Extracting Discriminating Features with Information Distance Measures

Next, we discuss how to leverage domain contexts to extract discriminating features automatically based on information theory. First, we define domain contexts as follows:

**Definition 1.** *Domain contexts* are a set of documents $C = c_1, c_2, ... c_n$ where each $c_i$ represents a specific topic of interest in the domain.

Conceptually, domain contexts represent the knowledge of a domain, which include a structure where each component of this structure is a document $c_i$. As an example, the course syllabi are domain contexts for the course forum. They contain high level knowledge of the course domain, and each week of the syllabus is a different document that focus on a specific topic of interest. Domain contexts are usually produced by domain experts that have high knowledge of the domain.

Next, we formally define Discriminating Features (DF) as follows:
Definition 2. Given a data set with domain contexts $C = \{c_1, c_2, \ldots, c_n\}$, two features $(f_i, f_j)$ are considered as **discriminating features** if the information distance between $(f_i, f_j)$ with respect to $C$ is larger than a threshold $\varepsilon$ (i.e. $\text{dist}(f_i, f_j) > \varepsilon$).

To evaluate the information distance between two features, the target variable at hand is the structure of the domain contexts $C$. We can measure the information distance between two features $f_i$ and $f_j$ as the distance between the distributions of $C$ they induce: $P(C|f_i)$ and $P(C|f_j)$.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig41.png}
\caption{In the epidemics forum data set, the probability distribution of syllabus structure $C$ induced by social, network, and vaccine.}
\end{figure}

An example of the distributions of $P(C|f_i)$ and $P(C|f_j)$ in data from the epidemic course syllabus is shown in figure 4.1. Consider the line for word *social*. The horizontal axis represents different (order-irrelevant) syllabus weeks. The vertical axis indicates the probability of each week given the word *social*. The shape of the line shows the probability distribution of $P(C|\text{social})$, which is mostly concentrated on week 4 and week 2.

We can easily observe in figure 4.1 that the distributions of *social* and *network* are more similar but they are very distinct from the distribution of *vaccine*. Conceptually,
social and network are more closely related as they indicate how diseases spread through the human social network. On the contrary, vaccine is a different topic in the epidemics domain and it is more relevant to how to prevent the disease.

This example expresses our intuition behind the task of finding discriminating features: the information distance between $f_i$ and $f_j$ can be measured by the difference of the probability distributions between $P(C|f_i)$ and $P(C|f_j)$. Now we turn to the question of how we quantify the difference between two probability distributions.

4.2.1 KL Divergence and KL Distance

Kullback-Leibler divergence (KL) \cite{32} is commonly used to quantify the difference between two probability distributions, which is defined as

$$\text{KL}(X||Y) = \sum_i p(x_i) \log \frac{p(x_i)}{p(y_i)}$$

We can interpret KL divergence as the expected bits of information required to code $X$ when using a code optimized to code $Y$. As a divergence measure, it satisfies the positive definiteness property, which implies that $\text{KL}(X||Y) \geq 0$, and $\text{KL}(X||Y) = 0$ when $X = Y$.

However, it is not difficult to see that the KL divergence is not symmetric. And because of this, it is not suitable to use it directly to measure the dissimilarities between
the words. We can transform the KL divergence into a symmetric form as **Kullback-Leibler distance** (KLD) \([32, 44]\), defined as

\[
KLD(X \| Y) = \frac{1}{2}(KL(X \| Y) + KL(Y \| X))
\]

The KL distance gives us a symmetric form of KL divergence. It retains the characteristics of KL divergence such as its positive definiteness. We can use this measure to calculate the word dissimilarity between \(w_i\) and \(w_j\) given domain contexts \(C\) by calculating \(KLD(p(C|w_i) || p(C|w_j))\), where \(p(C|w_i)\) is the conditional distribution of the domain context given the word \(w_i\). Since the discriminating features usually concentrate on different components in the domain contexts, which will result in a high KL distance score.

### 4.2.2 Loss of Mutual Information

In addition to KLD, we consider another metric called "**Loss of Mutual Information** (LMI)\([50]\)”, based on an intuition of maximizing the information loss when mixing the meaning of two words.

Let us first consider an example in the syllabus of the epidemics forum data set. The word *network* and *vaccine* has very different concentrations in the syllabus. *network* occurred frequently in week 4 while vaccine mostly occur in week 5. Conversely, these two words represent their own week in the syllabus since if we see the word *vaccine*, we will know that it is related to week 5’s content. Now imagine if we combine *network* and *vaccine* into a single feature \(f_m\), we then lose a key information to distinguish week 4
and week 5. As a matter of fact, the more discriminating or dissimilar the two words are, the more information it will lose due to the combination.

More formally, we can consider the random variable of word features $F = \{f_1, \ldots, f_i, \ldots, f_j, \ldots, f_v\}$ where each $f_i$ is a single word. We then measure the shared information between the domain contexts $C$ and the features $F$ as their mutual information $I(C;F)$. Then we pick two features $f_i$ and $f_j$ and combine them into a combined feature, $f_m = f_i, f_j$. The new word features becomes $F' = \{f_1, \ldots, f_m, \ldots, f_v\}$. We can measure the new mutual information after the combination, as $I(C;F')$. Loss of Mutual Information (LMI) is then captured by:

$$LMI(\{w_i, w_j\}, w_m) = I(C; W) - I(C; W')$$

We can further expand the above equation as

$$LMI = (p(w_i) + p(w_j))JS_{\pi_i, \pi_j}(p(C|w_i), p(C|w_j))$$

where $JS_{\pi_i, \pi_j}(p(C|w_i), p(C|w_j))$ is the weighted Jensen-Shannon divergence [36] between two probability distributions $p(C|w_i)$ and $p(C|w_j)$ with weights

$$\pi_i = \frac{p(w_i)}{p(w_i) + p(w_j)}; \pi_j = \frac{p(w_j)}{p(w_i) + p(w_j)}$$
Note that the weighted Jensen-Shannon Divergence is given by $JS_{\pi_i,\pi_j}(p_i,p_j) = \pi_i KL(\bar{p}||p_i) + \pi_j KL(\bar{p}||p_j)$, where $\bar{p} = \pi_ip_i + \pi_jp_j$. More details of the equations can be referred to [10].

LMI is effective to discover discriminating features since if we combine two discriminating from different $c_i$ into a single feature, the information loss will be greater than combing two similar words such as social and network. Also, LMI is strictly greater than zero [10].

As a brief summary, two information distance measures KL Distance and Loss of Information are introduced to measure the information distance between two words with respect to domain contexts $C$. They are both symmetric metrics and enable us to effectively extract discriminating features from domain contexts.
Chapter 5

Discriminating Features Sets

In this chapter, we introduce how to construct discriminating feature sets given the domain contexts using the information distance we introduced above.

First, we define **Discriminating Feature Set (DFS)** as follows:

**Definition 3.** A set of features \( F = \{f_1, f_2, ..., f_n\} \) is called a discriminating feature set if \( \forall (f_i, f_j) \in F, (f_i, f_j) \) are discriminating features.

There might be multiple discriminating feature sets \( S = \{F_1, F_2, ...F_n\} \) in the domain, where each \( F_i \) is a discriminating feature set.

To extract all discriminating feature sets, an obvious idea is to combine the most discriminating features greedily into sets using an agglomerative procedure, as shown in Algorithm 1.

![Diagram](image)

Fig. 5.1: Example of the first 5 iterations of algorithm 1 on the epidemics course forum data set
Algorithm 1 Constructing Discriminating Feature Sets

1: Input: Domain Context Corpus $C$
2: Output: Discriminating Feature Sets $S$
3: procedure CONSTRUCT DFS
4: Initialize $S$ as all features $S = \{ \{f_1\}, \{f_2\}, ..., \{f_n\} \}$
5: repeat
6:   $(i_{\text{max}}, j_{\text{max}}) = \text{argmax}_{i, j} \text{dist}(\{f_i\}, \{f_j\})$
7:   $\{f_m\} = \{f_i\} \cup \{f_j\}$
8:   $S = S \setminus \{\{f_{i_{\text{max}}}\}, \{f_{j_{\text{max}}}\}\} \cup \{f_m\}$
9: until after $k$ iterations or no more DF can be found
10: for all $s_i \in S$ do
11:    if $|s_i| == 1$ then
12:      $S \setminus s_i$
13:    end if
14: end for
15: end procedure

We will explain this algorithm with the epidemics course forum data set as an example, shown in figure 5.1. At first, all features are initialized as single words in the vocabulary. Then we calculate the information distance between all pairs of features in the syllabus and discovered that disease and vaccine has the largest information distance. We combine these two features into a single feature \{disease, vaccine\} and replace feature \{disease\} and \{vaccine\} with the new combined feature. In the next iteration, we found that the largest information distance between feature pairs are the new feature \{disease, vaccine\} and \{control\}, and we combine them into a three word feature. The similar process goes on and we discover a 4 word discriminating feature set after three iterations. Also, we found two discriminating feature sets (\{immunity, epidemic\} and \{pathogen, global\}) in the next two iterations. In total, after five rounds we derived three discriminating feature sets that contains 8 different words. It can continue to run for a certain number of iterations or until no more discriminating features can be found.
Using DFS is beneficial in discovering domain constraints such as cannot-links. We can encode a cannot-link relationship between each pair of features within each discriminating feature set. It also provides a clear view on the domain information such that a domain expert can easily evaluate the DFS result and fine tune the constraints.

Admittedly, there are some limitations of DFS. For example, consider three words *social*, *network*, and *vaccine*. We may find *network* and *vaccine* as a pair of discriminating features at the first step. But putting these two words in a DFS indicates that the program may not put *social* and *vaccine* into a same discriminating feature set since *social* and *network* are not discriminating. For this reason, it seems that we may have neglected the cannot-link constraint between *social* and *vaccine*. Fortunately, topic modeling algorithms provide a safe proof mechanism which will put *social* and *network* into the same topic due to their frequently co-occurrence. Since our program will still force *network* and *vaccine* into different topics, we can most of the time neglect the redundant constraint between *social* and *vaccine* to create more concise constraints.
Incorporating the Domain Constraints into Topic Models

In this chapter, we discuss how we can extend topic models to incorporate the discriminating feature sets that we discovered in the last chapter.

6.1 Dealing with Cannot-Links

First, we generate the cannot-links between each pair of words based on discriminating feature set. Suppose there are $k$ words in DFS $S$, we can derive $C^2_k$ cannot-links from the DFS by iterating all combinations. Next, we incorporate the cannot-links information into LDA in a similar way as the AMC model (topic modeling with Automatically generated Must-links and Cannot-links) [12]. AMC has the same graphical model as LDA, which is shown in Figure ??.

The generative process is simply as follows:

1. Choose $\theta_i \sim Dir(\alpha)$, where $i \in \{1, ..., M\}$ and $Dir(\alpha)$ is the Dirichlet Distribution for parameter $\alpha$.

2. Choose $\phi_k \sim Dir(\beta)$, where $k \in \{1, ..., K\}$

3. For each word $w_{i,j}$ where $i \in \{1, ..., M\}$ and $j \in \{1, ..., N\}$:
   a) Choose a topic $z_{i,j} \sim Multinomial(\theta_i)$.
   b) Choose a word $w_{i,j} \sim Multinomial(\phi_{z_{i,j}})$. 
Although sharing the same graphical model and generative process as LDA, we will employ a completely different parameter inference technique, which handles the constraints of cannot-links while sampling the topic for each word.

6.2 Sampling Technique

To incorporate cannot-links, we introduce a technique called "multi-generalized Polya urn" (M-GPU) model. The basic idea is if a word $w$ is assigned to a topic $t$, and if $w$ and $w_c$ share a cannot-link, then we will try to resample $w_c$ and force it to put in another topic $t_c$, where $t \neq t_c$. We will start by first introducing the basic Polya urn model and then generalize it to the M-GPU model.

6.2.1 Polya Urn Model

The Polya urn model[27] is a statistical model that describes the relationships between colored balls and urns. Originally in the Polya urn model, the urn contains $x$ red and $y$ green balls. Suppose we draw a ball from the urn randomly and record it color, we then add some extra ball to the urn based on the specific constraints. The question is, after multiple iterations, how will the colors and population in the urn evolve. This process usually relates to a self-reinforcing property usually known as "the rich get richer".

6.2.2 Simple Polya Urn Model (SPU)

In topic modeling, a term in the vocabulary can be viewed as a certain color of the ball, and a topic as a urn. The word distribution in a topic is the proportion of each
color in the urn. LDA can be viewed as a simple Pólya urn model (SPU) [20]. This process works as follows: when a ball is drawn from a urn with a particular color, the ball is replaced back to the urn along with an additional ball with the same color, as shown in Figure 1. This process corresponds to assigning a topic to a term in the Gibbs sampler of LDA. The more you see a term in a topic, the more you will assign it to this topic in the future, which is consistent with the "the rich get richer" property.

Fig. 6.1: Simple Pólya Urn Model

6.2.3 Generalized Pólya Urn Model (GPU)

Sometimes, the generalized Pólya urn model (GPU) [37, 42, 20] is useful. It differs from SPU in that: when drawing a ball of a particular color, two balls of that color is put back along with a certain number of balls with other colors. These addition balls of

\footnote{http://www.generationaldynamics.com/pg/ww2010.i.forecast090503.htm}
other added to the urn increase their probabilities in the urn. We call this a promotion of the other colored balls. This mechanism enables us to incorporate some constraints into the model such as must-links or synonyms.

### 6.2.4 Multi-generalized Pólya Urn Model (M-GPU)

The multi-generalized Pólya urn model (M-GPU) [12] is a further extension of GPU, which allows the interaction of multiple urns at a time. When a ball is randomly drawn, rather than returning two balls of the same color as in SPU, we may return a certain numbers of additional balls of each color to the urn, promoting their probabilities.

When incorporating cannot-link constraints, we will define two sets of urns in M-GPU. First, we define a set of topic urns $U^K_{d \in \{1...D\}}$, where each urn is for a document that contains the balls with $K$ colors (topics) and each ball has a color $k \in \{1...K\}$. This represents the document-topic distributions. The second set of urn is a set of term urns $U^W_{k \in \{1...K\}}$, where each urn is a topic containing terms $w \in \{1...V\}$ represented as the color of the balls. This corresponds to the topic-word distributions.

Due to the cannot-link constraint, two terms with a cannot-link relationship cannot both have a high probability in the same topic. When drawing a ball representing word $w$ from a term urn $U^W_k$, we want to transfer the balls that has a cannot-link with the terms $w$, say $w_c$, into other urns, which decreases the probability of $w_c$ of this topic while increase the probability of $w_c$ in another topic. Particularly, when choosing another urn for $w_c$, we want to transfer $w_c$ into a urn that make sense and put it into a urn that it should belong to. That is, we need to find a urn that has higher probability of $w_c$ than the current urn. To do this, we can randomly sample a urn from the urns
that have a higher probability of \( w_c \). However, it is possible that we cannot find other urns that has a higher probability of \( w_c \) and the current urn is the best choice of \( w_c \). In this case, considering the cannot-links are automatically generated, we might want to consider the possibility that this cannot-link may be incorrect, which indicates that we don’t want to put \( w_c \) elsewhere. It is possible that the data itself may strongly suggest these two terms should be in the same topic when they always appear together.

The M-GPU suggests us the technique to incorporate cannot-link constraints into the LDA, which is effectively implemented in the AMC model and used in this thesis. [12].
Chapter 7

Extracting Relationships between Derived Topics and Domain Contexts

Incorporating domain constraints into topic models is essentially to use domain contexts to influence and regulate the derived topics based on expert knowledge. For example, in the epidemics forum, the syllabus of the course specifically defined several topics of interests based on the course structures. Naturally, we would want to see what the discussions in the forum is about, corresponding to the course content, a.k.a. domain contexts. It is thus interesting to investigate the relationships between the derived topics and the original domain contexts.

In this chapter, we discuss the bipartite relationships between the derived topics and domain contexts. Then we propose to use Gibbs sampling to obtain the mappings between them. Finally, we discuss a possible direction of future research to apply it in information retrieval tasks.

7.1 Bipartite Graph Relationships

Conceptually, domain contexts are the high level texts generated by domain experts that provides an organized topics of the data. To some extent, it is like the domain ontology but it may not be comprehensive and precise. Some examples of the domain contexts include the syllabus of the course, types of cancer of a cancer forum, a list of all movies and their descriptions for a movie review forum, and so on.
If we view the topics in topic modeling as a thematic concept that is summarized by the top-k topic words, each domain context is then even more eligible to be view as a topic on its own because it contains high level descriptions that is semantically coherent. We may denote the derived topics learned from the large data as $t^D$, and the topics of domain contexts as $t^C$.

This problem can be formally defined as follows:

**Definition 4.** Given a set of documents corpus $D$ and their topics $t^D_{k=(1,...,K)}$ and a set of domain contexts $C$ containing topics of each domain context elements $t^C_{n=(1,...,N)} \in t^C$, we want to construct the bipartite relation $R$ between $T^D$ and $T^C$ where $R(t^D_k, t^C_n)$ indicate that the $k^{th}$ topic of the document corpus is semantically relevant to the $n^{th}$ domain context element.

This relationship can be effectively calculated by the Gibbs sampling technique.

### 7.2 Gibbs Sampling

Gibbs sampling is a popular technique for parameter inference in topic modeling [22]. Suppose we have a document corpus with $K$ topics, the probability of the $i^{th}$ word in document $d$ is:

$$P(w_i) = \sum_{j=1}^{K} P(w_i | z_i = j)P(z_i = j)$$

We can derive the posterior distribution of topic $z$ as:
\[ P(z|w) = \frac{P(w, z)}{\sum_i P(w, z)} \]

However, this distribution cannot be computed directly because the sum in the denominator is intractable. We then employ the Gibbs sampling, a Markov Chain Monte Carlo, to condition the current topic \( z_i \) on the current status \( z_{-i} \), which is the current topic assignment of all other words excluding the \( i^{th} \) word. We can derive that,

\[
P(z_i = j | z_{-i}, w) \propto \frac{n^{(w_i)}_{-i,j} + \beta}{n^{(i)}_{-i,j} + \beta \cdot K + \alpha} \]

where \( n^{(i)}_{-i,j} \) is the count that excludes the current assignment of \( z_i \), and \( V \) is the vocabulary size.

Based on the above equation, after deriving the topics \( t^D \), we can simply apply Gibbs sampling on domain context \( C \), and derive the topic assignment of all words in \( C \) and the document-topic distribution over each domain context element \( t^C \). We construct a link \( R(t^C_n, t^D_m) \) for each \( m \) in the top 3 topics of \( t^D \).

### 7.3 Future Extensions in Information Retrieval

With the relationship mappings between the derived topics and the domain contexts, it is possible for us to extend it into information retrieval tasks.

As an example, suppose the teachers of the online course are interested in whether the students have confusions or difficulties in the fifth week’s material, they can start
from the fifth syllabus and query the top 3 relevant topics of the forum data based on the relationship $R$ we learned from section 7.2. The fifth week may introduce several different topics, and the teachers can choose a particular topic of interest to further investigate. Given that topic models can also produce the most related document for each topic, the teachers can look into the actual discussion easily.

This approach has a very important advantage. Note that first of all, the teacher does not need to enter a specific keyword for the query and the system simply enable the users to browse the data that’s organized semantically. Second, the retrieved documents in the end may or may not have the same keyword as in the syllabus itself but their semantic are relevant, because they are linked in the topical space. This feature could be very useful sometimes for the teachers to learn the emerging problems that are not covered in the class but still related to the course material.
8.1 Epidemics Course Forum

Our first dataset contains all of the forum posts from the "Epidemics - the Dynamics of Infectious Diseases" course on Coursera.org. This course consists of eight weeks of materials and each week has a different topic of focus. The discussion forum that associates with this course is a convenient platform created by Coursera.org where students and professors can communicate. Many people ask questions on the discussion forum and get answered by other students or course staffs. In total, the whole forum contains 3,624 discussion threads, which has 38,021 posts and replies. In the forum structure aspect, this forum is divided into several sub-forums such as "General Discussions", "Weekly Lectures", and "Study Groups", etc. We found that the "Weekly Lectures" sub-forum, which contains eight sub-sub-forums for each week, has the richest domain specific information. Most discussions in the other forums are general chats that are not closely related to the epidemics topic of this course. In this thesis, we only extract the forum posts from the "Weekly Lectures" sub-forum to form a domain specific data that contains rich information. It has 1,942 threads and 29,274 posts. Our forum posts are short and each post is made of 24 words in average. We aggregate all the posts under the same thread to form a document to achieve better topic modeling results [26, 38].

1https://www.coursera.org/learn/epidemics
For the domain contexts data, we used the syllabus of this epidemics course. In each weeks of syllabus, we extracted the title, overview, objectives and a one sentence description for each course video as our domain contexts data. In total, there are 8 contextual documents and around 420 words per document.

8.2 Machine Learning Course Forum

Similar to the epidemics forum, we also obtained the data from the "Machine Learning" course from Coursera.org. Similar to Epidemics forum, the machine learning forum consists of the student discussions of the machine learning course. The syllabus of the course is separated into 11 weeks, indicating that it contains 11 domain context elements. The data set contains 2,936 threads, and we preprocess the thread and syllabus in the same way as the epidemics forum data.

8.3 Breast Cancer Forum

The third dataset is acquired from the online cancer forum maintained by American Cancer Society (ACS). This forum serves as an online cancer support community, known as the Cancer Survivors’ Network (CSN), where cancer patient, cancer survivors and their families can share information about cancer and their emotions. Many people seek information or emotion support from the community about cancer on this forum. We conduct our experiments on the Breast Cancer sub-forum which contains posts between June 2000 to June 2012. It contains 20,714 discussion threads and 247,271

\[^2\text{https://www.coursera.org/learn/machine-learning}\]
posts in total. We sampled 2,057 threads in random, which has 24,246 posts, to conduct our topic modeling experiments as the large dataset may crash the DF-LDA model [12].

To extract relevant domain contexts for the breast cancer forum, we used the "learn about breast cancer" page\(^3\) on ACS as it distinguishes different aspects of breast cancer such as cancer treatments and the cause of cancer. We aggregated similar information and obtained 6 documents with around 3,000 words per document.

### 8.4 Data Preprocessing

Here, we describe the preprocessing steps that we applied. First, we remove all non-standard word characters such as HTML tags, URLs, special characters, and numbers in the data. Second, we removed the words that seem to be typos filtered by a standard English dictionary\(^4\). Third, we lemmatized our corpus and removed the standard stop words. After the data preprocessing, there are 27,085 words in the dictionary and 711,797 word tokens in total.

\(^3\)http://www.cancer.org/cancer/breastcancer
\(^4\)CMUDict 0.7b developed by Carnegie Mellon University
Chapter 9

Experiments

In this chapter, we evaluate and compare the proposed domain constraints extraction methods using topic models as an application. We first compare the quality of domain constraints found by LMI and KLD. Second, we quantitatively evaluate the quality of the generated topics, using the topic coherence metric[42]. Third, we empirically study the topic quality by showing that LMI can effectively distinguish different concepts into different topics. Finally, we present the relationships between the derived topic and the domain contexts.

9.1 Evaluating Discriminating Feature Sets

In this thesis, we introduced KL Distance and Loss of Mutual Information in chapter 4 as our proposed information measures to discover the discriminating features. Then, we developed an agglomerative algorithm in chapter 5 to combine the discriminating features into discriminating feature sets.

Our first task here is to study whether the discovered discriminating feature sets can fit the expert’s expectation and can truly discover the discriminating features. We used our epidemics course forum data set for this task and we extracted the DFSs from the syllabus using KLD and LMI respectively.
Table 9.1: Keywords picked by domain experts from each weeks’ syllabus of epidemics course

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>microparasite</em></td>
<td>immunity</td>
<td>extinction</td>
<td>social</td>
</tr>
<tr>
<td><em>macroparasite</em></td>
<td>antibody</td>
<td>cycles</td>
<td>network</td>
</tr>
<tr>
<td><em>virus</em></td>
<td>antigen</td>
<td>seasonal</td>
<td>superspreader</td>
</tr>
<tr>
<td><em>bacteria</em></td>
<td>evolution</td>
<td>influenza</td>
<td>metapopulation</td>
</tr>
<tr>
<td><em>infection</em></td>
<td>resistance</td>
<td>diversity</td>
<td>transporation</td>
</tr>
<tr>
<td><em>disease</em></td>
<td>genetic</td>
<td>host</td>
<td>communication</td>
</tr>
<tr>
<td><em>microbiome</em></td>
<td>inflammation</td>
<td>emergence</td>
<td>surveillance</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>prevention</td>
<td>eradication</td>
<td>emergence</td>
<td>SARS</td>
</tr>
<tr>
<td>treatment</td>
<td>elimination</td>
<td></td>
<td>globalization</td>
</tr>
<tr>
<td>vaccine</td>
<td>campaigns</td>
<td>drug</td>
<td>incidence</td>
</tr>
<tr>
<td>measles</td>
<td>quarantine</td>
<td>antibiotic</td>
<td>prevalence</td>
</tr>
<tr>
<td>smallpox</td>
<td>behavior</td>
<td>prescription</td>
<td>biodiversity</td>
</tr>
<tr>
<td>herd</td>
<td>avoidance</td>
<td>virulence</td>
<td>dilution</td>
</tr>
<tr>
<td>clustering</td>
<td>mosquito</td>
<td>selection</td>
<td>zombie</td>
</tr>
</tbody>
</table>

Table 9.1 shows the top-7 keywords of each weeks’ syllabus content hand-picked by domain experts. There are eight weeks of the syllabus, and the topic (title of the syllabus) of each week is labeled with their id.

We present the discriminating features learned with KLD and LMI with a color code. A different color indicates that the word is belong to different weeks as in table 9.1. The more colorful a set is, the better is the quality of the discriminating feature set. We present the result in Figure 9.2, where each row is a discriminating feature set.

We evaluate the quality of the discriminating feature sets based on the following measures:
Table 9.2: Discriminating Feature Sets learned with KLD and LMI in epidemics course syllabus

<table>
<thead>
<tr>
<th></th>
<th>KLD</th>
<th>LMI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>infectious, common, final, space, infection</td>
<td>reproductive, disease, organism, control, drug,</td>
</tr>
<tr>
<td></td>
<td>problem, intervention, impact, extinction, immune</td>
<td>resistance, network, host, immune, vaccine</td>
</tr>
<tr>
<td></td>
<td>return, approach, sic, human</td>
<td>cause, human</td>
</tr>
<tr>
<td></td>
<td>particularly, population, lecture, determine</td>
<td>antibiotic, surveillance, person, system, impact</td>
</tr>
<tr>
<td></td>
<td>around, control, third, ecological, vaccine</td>
<td>infectious, herd</td>
</tr>
<tr>
<td></td>
<td>prevent, health, connect, take</td>
<td>infect, intervention</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pathogen, social</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pathogens, global</td>
</tr>
<tr>
<td></td>
<td></td>
<td>immunity, epidemic, evolution</td>
</tr>
</tbody>
</table>

\[
Consistency = \frac{\text{number of words in the expert’s selection}}{\text{total words of DFS}}
\]

\[
Diversity = \frac{\text{number of words from different domain context elements}}{\text{total number of domain context elements}}
\]

The average consistency and diversity of the discriminating feature sets weighted by the size of DFS is summarized in table 9.3.

We have the following observations:

1. The consistency and diversity of the discriminating feature sets discovered by LMI are better than KLD. Although the DFS discovered LMI is still not fully diversified,
Table 9.3: Weighted Average Consistency and Diversity of the discriminating feature sets

<table>
<thead>
<tr>
<th></th>
<th>Consistency</th>
<th>Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>KLD</td>
<td>29.63%</td>
<td>15.74%</td>
</tr>
<tr>
<td>LMI</td>
<td>64.29%</td>
<td>38.84%</td>
</tr>
</tbody>
</table>

it is still very useful. For example, the DFS \{infection, herd\} suggests to put the these two key words into different topics, which will serve as magnets to attract other topic words and results in better topics. Note that in the first DFS of LMI, it contains 6 different colors, indicating that it can discriminate 6 week’s content. The other two weeks’ keyword are covered in the third and seventh DFS by contrasting them with fourth and first week’s keywords respectively. In comparison, KLD can only identify keywords from 5 weeks (5 colors), and can only discriminate 4 of them since if the DFS only has one color, it does not have significant discriminating power to separate two different week’s content.

2. In KLD, there are a lot of domain independent words such as "space", "particularly", and "human", which carries little information values. In LMI, although there are also a few uncolored words, but in the epidemics domain, they are more meaningful words such as "organism", "drug", "epidemic".

This result indicate that the DFS generated by LMI are more correlated with the constraints provided by domain experts.
9.2 Topic Coherence

This section investigates the effectiveness of the extracted discriminating feature sets when incorporated into topic models. Traditionally, the quality of topics is commonly evaluated with perplexity\cite{54}, but Chang et al.\cite{11} found that perplexity does not always correlate with semantically coherent topics. It is usually inconsistent with human judgements. Because of this, we choose to compare the quality of the topics with topic coherence measure, developed by Mimno et al.\cite{42}. Topic coherence measure is shown to correlate well with human labeling\cite{51}, and serve as a popular topic model evaluation methods in recent years\cite{12, 14}. The topic coherence for each topic is defined as

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_{m}^{(t)}, v_{l}^{(t)}) + 1}{D(v_{l}^{(t)})},$$

where \(M\) is the number of topic words we define in the topic word distribution, and \(D(v_{m}^{(t)}, v_{l}^{(t)})\) is the total number of documents containing both \(v_{m}^{(t)}\) and \(v_{l}^{(t)}\).

We then train our topic model by extending the M-GPU model implementation\footnote{https://github.com/czyuan/AMC} developed by Chen et al.\cite{12} and incorporate the discriminating feature sets. Two baseline models are compared: 1) the LDA model\cite{8} in which no domain knowledge is included; and 2) the GK-LDA model\cite{14} where general knowledge such as antonyms are incorporated. We sample the posterior latent variables for 500 iterations in total (first 100 iterations for burn-in). The Dirichlet parameters are set to \(\alpha = 1, \beta = 0.1\). The number of topics \(k\) is set to be \(k = \{15, 20, 25, 30, 35, 40, 45, 50\}\) in all of our data set.
And the number of domain contexts is $|C| = 8$ for the epidemics course forum, $|C| = 6$ for the breast cancer forum, and $|C| = 11$ for the machine learning course forum. We extracted the discriminating features sets with KLD and LMI metrics with 20 iterations in algorithm 1. Finally, we compute the topic coherence with the top 30 topic words in the derived topics. The results is shown in Figure 9.1.
Fig. 9.1: Topic Coherence scores on all datasets using topic number $k = 15 \sim 50$
Note that topic coherence is negative because it is the log of a number between 0 \sim 1, and a higher value indicates better topic quality. From Figure 9.1, we can observe the following:

1. In the epidemics course forum and breast cancer forum data set, both of our models (MGPU+LMI and MGPU+KLD) can significantly \((p < 0.01\) on paired sample student t-test based on different number of topics) out-perform the baseline LDA and GK-LDA when incorporating domain constraints, no matter whether we are using KLD or LMI as the information metric. However, the difference between KLD and LMI is not as significant as compared to LDA.

2. In the first two data sets, when the number of topics is large, our methods (MGPU+LMI and MGPU+KLD) performs even better compared to the baselines. We believe that when given more topics, the discriminating features we identify will have more freedom to be correctly separated into different topics, which can further increase topic coherence.

3. In machine learning forum data set, the difference between LDA and MGPU-LMI and MGPU-KLD is not as obvious as that in the previous data sets. Also, we can observe that GK-LDA can significantly out-perform other methods. Our hypothesis of this phenomenon is that the machine learning course forum is more related to general discussions and less focused on the domain keywords. Although the syllabus contains standard machine learning topics such as "linear regression" or "neural networks", the forum discussion has a lot of topics discussing the details of coding and implementations of the machine learning algorithms. For example, the resulting topics contains a lot of detailed keywords such as "matrix", "theta" or "error", which are not covered by the
syllabus (the topics and DFS of machine learning forum is shown in appendix). This indicates that our method could have the potential to judge whether the domain context fit the data set closely.

In summary, we can see that the automatic extracted domain constraints can certainly help to produce more coherent topics when incorporating with the well-fit knowledge.

9.3 Relationships between Derived Topics and Domain Contexts

In this section, we evaluate the relationships between the derived topics and domain contexts discussed in Chapter 7. In the epidemics forum, we ran LMI incorporated AMC model with 20 topics in the epidemics forum and we extracted the top 5 relevant topics for each syllabus weeks to form the links. We present the topics as a bipartite graph in figure 9.2 if they have 1-4 links. We found that the topics that have more than 4 links are usually general topics and those that has 0 links are usually emerging topics, which are shown in table 9.4.
Fig. 9.2: Relationships between derived topics and domain contexts in the epidemics course forum dataset

Table 9.4: General topics and emerging topics in epidemics course forum

<table>
<thead>
<tr>
<th>General Topics</th>
<th>Emerging Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>disease people</td>
<td>flu people</td>
</tr>
<tr>
<td>transmission</td>
<td>infection</td>
</tr>
<tr>
<td>public</td>
<td>disease</td>
</tr>
<tr>
<td>country</td>
<td>number</td>
</tr>
<tr>
<td>many</td>
<td>individual</td>
</tr>
<tr>
<td>issue</td>
<td>population</td>
</tr>
<tr>
<td>think</td>
<td>transmission</td>
</tr>
<tr>
<td>important</td>
<td>infectious</td>
</tr>
<tr>
<td>world</td>
<td>pathogen</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We can have the following observations from figure 9.2 and table 9.4.

1. Some topics are very dedicated to one particular domain context element. For example, topic 2 is only related to syllabus 2; and topic 6 is only related to syllabus 5.
Judging from the topic words, their top few words are also highly consistent with the syllabus content.

2. Looking from the derived topics side, some topics are like super topics that contains the meanings of multiple syllabus contexts. Such examples include topic 4 and topic 7. However, although they are linked to multiple syllabus contexts, their topic words are still quite coherent. Topic 4 is mostly related to pathogen, and the change and mutation of pathogens, which is more relevant to the 2nd and 3rd week’s course.

3. We can see that the mappings between syllabus and derived topics relatively local. For example, topic 1 is related to syllabus 1 and 2; topic 3 is related to syllabus 3 and 6; and topic 8 is related to syllabus 7 and 8. This indicates that the course materials are coherently related. Each weeks of the course is very likely related to the contents of last week semantically, and the discussions in the forum are also very likely to correspond to multiple relevant syllabus contexts.

4. From table 9.4, we can identify seven emerging topics such as washing hands, scientific research, particular diseases, medical treatments, and online course. These topics are very relevant to the epidemics domain but they are not directly taught in the online course. This finding provides a useful topical structure for the lecturers of the course to understand the topics that the students are interested in and can guide them to learn more about the domain.
Chapter 10

Conclusions

This thesis proposes an automated domain constraints extraction method to learn prior knowledge from domain contexts. We first introduce two information distance measures, namely Kullback-Leibler distance and Loss of Mutual Information, to discover the distance with respect to the domain contexts of each pair of features. We then propose an agglomerative algorithms to leverage the information distance to construct discriminating feature sets. Furthermore, we employ a knowledge-based topic model based on Multi-generalized Pólya Urn Model to incorporate the domain constraints. And finally, we propose to use Gibbs sampling technique to construct the bipartite relationships between the derived topics and the structured domain knowledge. Experimental results using epidemics course forum, breast cancer forum and the machine learning forum show that the automatically extracted domain constraints are useful to generate significantly better topic quality.

10.1 Future Work

In our future work, we may first study how we can improve the extracted prior knowledge. We can investigate more contextual prior knowledge and compare what data can be better leveraged to improve the topic qualities. We may also consider other techniques to calculate the information distance. For example, we may use Word2Vec
to calculate the minimum cosine similarities between word vectors to discover discriminating features. Second, we would like to conduct more comprehensive quantitative and qualitative experiments to compare with other topic discovery techniques such as using Seeded LDA, and deep learning approaches [33] which attract a lot of attentions lately. Third, further investigating why machine learning forum performed differently compared to the other two data sets may also help us to gain a deeper understanding about the domain constraints. Finally, extending our findings in the bipartite relationships between domain contexts and the derived topics to improve information retrieval can also show its potential.
Appendix

DFS and topics of machine learning data set

In chapter 9.2, we observed that in machine learning data set, our discriminating feature set cannot help to improve topic qualities. Our hypothesis is that the topics in the forum discussions are not correlate with the knowledge mined from the syllabus well. In the course syllabus, it discussed about different types of machine learning methods while in the forum topics, most topics are focused on the implementation details of the machine learning algorithms. The DFS is shown in table A.1 and the actual topics on 20 topics with LMI is shown in table A.2.

Table A.1: Discriminating Feature Sets learned with KLD and LMI in machine learning course syllabus

<table>
<thead>
<tr>
<th>KLD</th>
<th>LMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>detection, introduction, week, unsupervised</td>
<td>network, regression, algorithm</td>
</tr>
<tr>
<td>representation, core, help</td>
<td>model, understand</td>
</tr>
<tr>
<td>ocr, machine, idea, inspire</td>
<td>algebra, help, system</td>
</tr>
<tr>
<td>vector, classify, computer, use</td>
<td>application, data, assignment</td>
</tr>
<tr>
<td>concept, support, anomaly, learn, extend, logistic, photo, congratulations, classifier, scale</td>
<td>introduce, work</td>
</tr>
<tr>
<td>train, use, performance</td>
<td>one, need</td>
</tr>
<tr>
<td>multiple, example</td>
<td>regularization, machine, linear, learn, logistic, neural, recommender, octave, product</td>
</tr>
</tbody>
</table>
Table A.2: 20 topics in machine learning course forum discovered by the M-GPU+LMI model

<table>
<thead>
<tr>
<th>study</th>
<th>regression</th>
<th>machine learning</th>
<th>N/A</th>
<th>vectors</th>
<th>octave</th>
<th>data</th>
<th>parameters</th>
<th>training</th>
<th>plot</th>
</tr>
</thead>
<tbody>
<tr>
<td>group</td>
<td>regression</td>
<td>learn</td>
<td>not</td>
<td>you</td>
<td>octave</td>
<td>theta</td>
<td>train</td>
<td>set</td>
<td>plot</td>
</tr>
<tr>
<td>interest</td>
<td>linear</td>
<td>machine</td>
<td>need</td>
<td>vector</td>
<td>matlab</td>
<td>feature</td>
<td>theta</td>
<td>cost</td>
<td>data</td>
</tr>
<tr>
<td>study</td>
<td>logistic</td>
<td>course</td>
<td>just</td>
<td>theta</td>
<td>use</td>
<td>version</td>
<td>model</td>
<td>error</td>
<td>lambda</td>
</tr>
<tr>
<td>work</td>
<td>function</td>
<td>thank</td>
<td>get</td>
<td>column</td>
<td>use</td>
<td>use</td>
<td>theta</td>
<td>value</td>
<td>test</td>
</tr>
<tr>
<td>hello</td>
<td>use</td>
<td>algorithm</td>
<td>think</td>
<td>dimension</td>
<td>high</td>
<td>value</td>
<td>test</td>
<td>validation</td>
<td>pca</td>
</tr>
<tr>
<td>like</td>
<td>class</td>
<td>find</td>
<td>work</td>
<td>row</td>
<td>number</td>
<td>derivative</td>
<td>partial</td>
<td>iteration</td>
<td>problem</td>
</tr>
<tr>
<td>please</td>
<td>predict</td>
<td>book</td>
<td>make</td>
<td>sun</td>
<td>run</td>
<td>dataset</td>
<td>point</td>
<td>cross</td>
<td>use</td>
</tr>
<tr>
<td>everyone</td>
<td>problem</td>
<td>conjurer</td>
<td>try</td>
<td>example</td>
<td>file</td>
<td>fit</td>
<td>link</td>
<td>algorithm</td>
<td>graph</td>
</tr>
<tr>
<td>add</td>
<td>sigmoid</td>
<td>good</td>
<td>value</td>
<td>will</td>
<td>one</td>
<td>one</td>
<td>one</td>
<td>point</td>
<td>cross</td>
</tr>
<tr>
<td>guy</td>
<td>value</td>
<td>take</td>
<td>will</td>
<td>loop</td>
<td>link</td>
<td>link</td>
<td>link</td>
<td>square</td>
<td>curve</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>gradient</th>
<th>neural nets</th>
<th>code</th>
<th>homework</th>
<th>lecture</th>
<th>parameters</th>
<th>Q&amp;A</th>
<th>parameter</th>
<th>code</th>
<th>video</th>
</tr>
</thead>
<tbody>
<tr>
<td>descent</td>
<td>network</td>
<td>code</td>
<td>error</td>
<td>course</td>
<td>theta</td>
<td>question</td>
<td>use</td>
<td>code</td>
<td>video</td>
</tr>
<tr>
<td>theta</td>
<td>neural</td>
<td>function</td>
<td>submit</td>
<td>week</td>
<td>cost</td>
<td>answer</td>
<td>sigma</td>
<td>code</td>
<td>video</td>
</tr>
<tr>
<td>layer</td>
<td>test</td>
<td>file</td>
<td>exercise</td>
<td>use</td>
<td>one</td>
<td>function</td>
<td>time</td>
<td>detail</td>
<td>url</td>
</tr>
<tr>
<td>alpha</td>
<td>output</td>
<td>use</td>
<td>use</td>
<td>programme</td>
<td>get</td>
<td>algorithm</td>
<td>value</td>
<td>one</td>
<td>url</td>
</tr>
<tr>
<td>use</td>
<td>input</td>
<td>get</td>
<td>line</td>
<td>time</td>
<td>thank</td>
<td>system</td>
<td>loop</td>
<td>matlab</td>
<td>browser</td>
</tr>
<tr>
<td>equation</td>
<td>hide</td>
<td>port</td>
<td>try</td>
<td>lecture</td>
<td>gradient</td>
<td>use</td>
<td>thank</td>
<td>language</td>
<td>additional</td>
</tr>
<tr>
<td>value</td>
<td>train</td>
<td>thank</td>
<td>work</td>
<td>assignment</td>
<td>value</td>
<td>give</td>
<td>get</td>
<td>like</td>
<td>page</td>
</tr>
<tr>
<td>normal</td>
<td>unit</td>
<td>answer</td>
<td>exercise</td>
<td>thank</td>
<td>function</td>
<td>example</td>
<td>error</td>
<td>python</td>
<td>question</td>
</tr>
<tr>
<td>get</td>
<td>example</td>
<td>case</td>
<td>submission</td>
<td>complete</td>
<td>regularization</td>
<td>correct</td>
<td>find</td>
<td>data</td>
<td>thank</td>
</tr>
<tr>
<td>iterations</td>
<td>number</td>
<td>value</td>
<td>assignment</td>
<td>get</td>
<td>grad</td>
<td>user</td>
<td>index</td>
<td>better</td>
<td>chrome</td>
</tr>
</tbody>
</table>
Bibliography


