Mining Information from Heterogeneous Sources: A Topic Modeling Approach [*]

Rumi Ghosh
Social Computing Lab
Hewlett Packard Research Labs
Palo Alto, California, USA
rumi.ghosh@hp.com

Sitaram Asur
Social Computing Lab
Hewlett Packard Research Labs
Palo Alto, California, USA
sitaram.asur@hp.com

ABSTRACT

In recent years, the phenomenal growth and popularity of social media, news and discussion websites has led to a vast number of information sources available online. These sources generate massive amounts of real-time content on a daily basis making it increasingly difficult to glean true and useful information from them. Automatically categorizing and compressing important contextual information from these sources is crucial for tasks such as web document classification and summarization. Therefore, in this paper, we propose a novel topic modeling framework- *Probabilistic Source LDA* which is designed to handle heterogeneous sources. *Probabilistic Source LDA* can compute latent topics for each source, maintain topic-topic correspondence between sources and yet retain the distinct identity of each individual source. Therefore, it helps to mine and organize correlated information from many different sources. At the same time, it aids in automatically reducing noise and redundancy in the information gathered. Using real data on the US elections 2012, we demonstrate that our *Probabilistic Source LDA* method can extract highly relevant latent topics while maintaining topic-topic congruence between different sources.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval—*Information filtering*; H.3.3 [**Information Systems**]: Information Search and Retrieval—*Retrieval Models*

General Terms

Experimentation, Measurement, Method

Keywords

Heterogeneous Sources, Integrating Sources, Topic Models, Data Mining

1. INTRODUCTION

Information sources on the Web have grown rapidly in recent years. This proliferation of social media, news, blogs and discussion groups has ushered in an era of information overload. It is hard to discount any of these sources of information, and yet, there is a vast amount of redundancy and noise that has to be overcome in order to extract relevant and actionable information from them. Additionally, these sources generate content that differ in size, frequency, expertise and relevance. For example, twitter produces tweets which are short and are posted at high rates. However they are known to contain high degrees of noise in them [1]. News articles are longer than tweets and are written to cover particular events. It is common for several news stories to feature the same event. Blogs are typically longer and less frequent than news or tweets. They also tend to be more opinionated and subjective. Applications like social search and document recommendation require web data to be suitably crawled and categorized for easy retrieval. Given the different varieties of sources, the task of integrating and understanding information from these heterogeneous sources is both a difficult and important challenge.

Topic models such as LDA [2] and PLSA [6] have been developed and used extensively for document analysis. However, while their use with text documents has been successful, their performance is less impressive with social media feeds such as Twitter, which are smaller and noisier than documents. Furthermore, there has not been extensive research on the application of these methods in the context of multiple heterogeneous information sources. The only existing research that we are aware of has been on comparative mining of 2 to 3 sources [7, 14]. Our goal is to aggregate multiple heterogeneous feeds (documents) of web data to extract global topics from them, while also *preserving the essential inter-document, intra-document, intra-source and inter-source statistical relationships* that are essential for basic tasks like classification. We wish to consider sources as diverse as news, radio broadcasts, discussions and social media and discover topics with maximum *coverage i.e.* that capture the main essence or information in each of these feeds in different sources. Finally we aim to ensure that there is *topic-topic correspondence* between sources.

One possible way to combine heterogeneous sources would be to naively merge documents from all sources into a single collection and subsequently apply existing topic modeling methods like LDA on it. However, this has several

[*]MDS'13, August 11 2013, Chicago, IL, USA Copyright 2013 ACM 978-1-4503-2321-5/13/08 ... $15.00.

drawbacks. First of all, the results might be biased in favor of a source with a lot of documents. Also, it would fail to consider the heterogeneity present in the sources. Such a process would not help us to discover the distinct topical characteristics of each individual source. An alternative option would be to run existing topic models like LDA on each source individually. This would preserve the source-specific properties but would make it very difficult to obtain the topic-topic correspondence between different sources and the knowledge of the global topic distribution. Some authors have suggested finding common topics for all sources and distinct topics for individual sources as a possible way to combine sources [7, 14]. However, their method of extracting common topics is essentially equivalent to taking documents from all sources as a single collection which does not ensure that the distinct properties of the individual sources would be retained. Also, there would be no congruence between topics from different sources.

To solve the aforementioned problems, we develop a probabilistic extension to LDA - *ProbLDA* that can enable it to handle multiple heterogeneous sources by assuming a probability distribution across the sources. We present three methods by which weights can be assigned to sources - *a priori* Source LDA where the weights are chosen *a priori*, Primary Source LDA where one of the sources is fixed as the primary source and Dirichlet Source LDA where the sources are drawn from a Dirichlet Distribution. The resulting methods generate local and global topics, while preserving the topic-topic correspondence across them.

We conduct extensive experiments on aggregating real data pulled from various sources such as political blogs, radio broadcasts, social media and news all focusing on the 2012 US elections. We extract relevant topics for all the sources using all the methods that we have developed. Our experiments show that the proposed *ProbLDA* methods can generate topics that leverage the properties of each of the individual sources, while preserving topic-topic correspondence across sources, and between the global and local topics. Furthermore, we observed that the *Primary Source LDA* method was the best of the *ProbLDA* methods and gave superior performance to LDA when evaluated on the local topics detected.

To summarize, the contributions of this paper are as follows:

- A framework for mining correlated information from heterogeneous web sources in an integrated fashion.
- A novel topic modeling technique- *Probabilistic Source LDA (ProbLDA)* that can handle heterogeneous sources while preserving topic-topic correspondence.
- Experimental evaluation of the proposed methods using real-data from various different sources, showing that *ProbLDA* is very effective in obtaining relevant global topics, while preserving the inter-document, intra-document, inter-source and intra-source relationships.

Table 1 summarizes the notations we use in the paper. Throughout the paper, we use the general term *document* or *feed* to cover all the basic text collections. For example, in the context of news media, a feed would be a news article while for Twitter, a feed would represent a tweet. We refer to a word occurrence as a *token.*

K	number of topics
V	number of words in vocabulary
S	number of sources
M^s	number of documents in source s
$N^s_{d=1\cdots M^s}$	number of words in document D in source s
N^s	$N^s = \sum_{d=1}^{M^s} N^s_d$
N	$N = \sum_{s=1}^{S} N^s$
$\theta^s_{j=1\cdots M^s,k=1\cdots K}$	probability of topic k occurring in document j in source s
$\theta^s_{j=1\cdots M^s}$	topic distribution of document j in source s
θ^s	M^s dimensional vector where $\theta^s[j] = \theta^s_j$
θ	S dimensional vector where $\theta[s] = \theta^s$
$\phi^s_{k=1\cdots K,r=1\cdots V}$	probability of word r occurring in topic k in source s
$\phi^s_{k=1\cdots K}$	word distribution of topic k in source s
ϕ	S dimensional vector where $\phi[s] = \phi^s$
$\bar{\alpha}$	K-dimensional vector where $\bar{\alpha}[k] = \alpha_k$, the parameter for the dirichlet prior on the per-document topic distributions
$\bar{\beta}$	V-dimensional vector where $\bar{\beta}[r] = \beta_r$, the parameter for the dirichlet prior on the per-topic word distributions
$\bar{\gamma}$	probability distribution on sources
w^s_{ij}	j^{th} token in the i^{th} document in source s
Z^s_{ij}	topic of the j^{th} token in the i^{th} document in source s
C^s_{ij}	source generating the topic distribution of the j^{th} token in the i^{th} document in source s
$\bar{\mu}$	Dirichlet prior on source distribution

Table 1: Table of symbols

In the next section, we survey related work in the area. We describe Probabilistic LDA in Section 3. Section 4 contains a description of the datasets we created. Section 5 details our experiments and results. We conclude the paper with discussion and future work in Section 6.

2. RELATED WORK

We first discuss prior research on discovering topics of documents from a single source. Apart from topic modeling, alternative approaches for finding short descriptions of documents include *tf-idf*, latent semantic indexing (LSI) and PLSA. The *tf-idf* scheme [10] reduces each document to a vector of real numbers based on the inter-document and intra-document frequency of words. While it can be effective in identifying words to distinguish documents and can reduce an arbitrary-length document to a fixed length list of numbers, it provides a relatively small reduction in description length and does not tell us a lot about inter-document or intra-document relationships.

Latent semantic indexing (LSI) [3] improves upon the *tf-idf* reduction scheme by reducing the dimension of the space of *tf-idf* features to get a linear subspace capturing the most variance in the collection. Hofmann[6] proposed an alternative to this, the *probabilistic latent semantic indexing*(PLSA) technique, which is a more principled statistical approach based on mixture decomposition derived from a latent class model. However, PLSA depends on the probability of a document and since there is no natural way to assign probability to these documents, PLSA is not a well-defined generative model. Other problems with this approach include the number of parameters to be estimated growing linearly with the number of documents[2] and its tendency to over-fit data [7]. To overcome these shortcomings, Blei and others [2] proposed *latent Dirichlet allocation (LDA)*, a generative probabilistic model for modeling data or documents from a single source. In this paper we propose *Probabilistic Source LDA (ProbLDA)*, a topic modeling technique that explicitly takes the heterogeneity of different information sources into account.

Heterogeneous sources of information on the web like social media and traditional news sources have been studied in the past. However, most of the prior research on topic categorization has been limited to analyzing a single source. There has been research on extracting common topics from multiple streams of text [4, 12, 13], and extending PLSA and LDA approaches to extract local and shared topics from multiple sources [14, 7]. However, most of these works have focused on modeling only two or three different sources of data. We on the other hand detect latent topics from an ensemble of sources ranging from news sources, to social media websites, to blogs ands radio broadcasts and so on. The advantages of using LDA over PLSA have been described in the previous paragraph. Similarly [7] models text stream from two news sources which finds local topics and shared topics. A word belongs to either the local topic or the shared topic, the probability of which is drawn from a bernoulli distribution. While [4, 12, 13] detect only common or shared topics, [14] detects both shared and local topics. Like [14] modifying PLSA, [7] extended LDA to find the local and shared topics for correlated sources. Besides, in the work by Hong and others, [7], local topics are computed by using LDA on each source individually. The shared topics are detected by pooling in all documents from all sources into one big collection of documents. This big collection of documents then becomes a single source and topic modeling methods like LDA designed for single sources are used. We note that in existing techniques detecting both local and shared topics [7, 14], during shared(common) topic detection, the individual properties of each stream of data are not preserved. Furthermore, there is no correspondence between local topics of different sources. Nor is there correspondence between the local topics and the shared or global topics. To the best of our knowledge, ours is the first attempt to detect topics leading to automatic topic-topic correspondence for different local sources as well as between the global and local topics.

Unlike [5] where words are drawn from different vocabulary distributions pertaining to different sub-stories (number of sub-stories fixed by the user *a priori*) which are then used for summarization, in our case the different vocabulary distributions pertain to the actual sources. In contrast to the work on multilingual probabilistic topic models [11], our proposed method is generative and does not require a training set for determining topic-topic correspondence between different sources. We also note that our proposed model is very different from existing extensions of LDA such as DiscLDA [8] and others[9] that take additional labeling information into account.

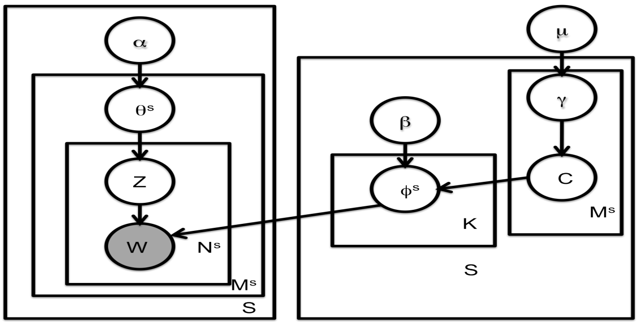


Figure 1: Plate notation of Probabilistic Source LDA (cf. Table 1)

3. PROBABILISTIC SOURCE LDA

In LDA, the probability of topic assignment given the document $P(Z^s_{j,x}|\theta^s_j)$ captures the intra-document relationship and the probability of word given topic $P(W^s_{j,x}|\phi_{Z^s_{j,x}})$ encodes the inter-document relationship within a collection of documents. The probability function which is maximized to obtain latent topics is a combination of *local* (intra-document) and *global* (inter-document) factors. Like LDA, the aim of most latent topic models for single sources is to capture the *local* (intra-document) and *global* (inter-document) relationships and at the same time reduce the description length of the documents. We extend this notion to documents from multiple sources. The goal then is to capture not only the inter-document and intra-document relationships but also the inter-source and intra-source relationships. Additionally, we aim to retain the distinguishable and essential statistical properties of each source. We propose a novel topic modeling method– *Probabilistic Source LDA (ProbLDA)* for this purpose. It helps us to detect topics across heterogeneous sources while maintaining the distinct characteristics of each individual source. For each source, it provides local topics while ensuring that there is topic-topic congruence between different sources. In other words, any local topic i of any source j would correspond to topic i of another source k. At the same time the characteristics of the local topics would conform to the properties of the local source. As an illustration, in the above case, the word distribution of local topic i for source j might be different from the word distribution of the local topic i corresponding to source k. The properties of global topics is determined from the properties of all the corresponding local topics. In this example, the properties of global topic i would depend on both the properties of local topic i in source j and local topic i in source k. This property is especially useful for automatic categorization and summarization of large web sources and can be used for social search.

Our proposed method maintains topic to topic correspondence by keeping the local relationship intact, but changing the global factor to capture not only the inter-document but also the inter-source relationships. It enables us to maintain the uniqueness of each individual source and at the same time lead to the global integration of their content.

We make the global factor, which generates the probability of a word given topic, conditional on the choice of the source from which this topic is drawn. The choice of the source depends on the distribution of sources $\bar{\gamma}$ which can be chosen *a-priori* or might be learned from the model.

Figure 1 gives the plate notation of the proposed model.

The Probabilistic Source LDA Model (*ProbLDA*):

1. Choose $\theta_i^s \sim Dir(\bar{\alpha})$ where $s \in \{1, 2, \cdots, S\}$, $i \in \{1, 2, \cdots, M_s\}$ and $Dir(\bar{\alpha})$ is the Dirichlet Distribution of $\bar{\alpha}$.
2. Choose $\phi_k^y \sim Dir(\bar{\beta})$ where $k \in \{1, 2, \cdots, K\}$ and $y \in \{1, 2, \cdots, S\}$.
3. For each word w_{ij}^s
   (a) Choose a topic Z_{ij}^s from Multinomial(θ_i^s)
   (b) Choose a source C_{ij}^s from Multinomial($\bar{\gamma}$)
   (c) Choose a word w_{ij}^s from Multinomial($\phi^{C_{ij}^s}{}_{Z_{ij}^s}$)

The w_{ij}^s are the only observable variables.

The total probability of the model is:

$$P(W, Z, C, \theta, \phi; \bar{\alpha}, \bar{\beta}, \bar{\gamma}) = \Pi_{s=1}^{S} \Pi_{i=1}^{K} P(\phi_i^s; \bar{\beta})$$
$$\Pi_{j=1}^{M^s} P(\theta_j^s; \bar{\alpha}) \Pi_{x=1}^{N^s} P(Z_{jx}^s | \theta_j^s) P(C_{jx}^s | \bar{\gamma}) P(W_{jx}^s | \phi_{{Z_{jx}}^{C_{x,j}^s}})$$

Using collapsed Gibbs sampling for inference, it can be shown that

$$\begin{aligned} &P(Z_{m,n}^s = k, C_{m,n}^s = y | Z_{-(m,n,s)}^s, C_{-(m,n,s)}^s, W; \bar{\alpha}, \bar{\beta}, \bar{\gamma}) \\ &\propto P(Z_{m,n}^s = k, Z_{-(m,n,s)}^s, C_{m,n}^s = y, C_{-(m,n,s)}^s, W; \bar{\alpha}, \bar{\beta}, \bar{\gamma}) \\ &\propto \frac{(n_{m,(.)}^{k,(s),-(m,n,s)} + \alpha_k)}{(\sum_{i=1}^{K} n_{m,(.)}^{i,(s),-(m,n,s)} + \alpha_i)} . \gamma_y . \frac{(n_{(.),v}^{k,(y),-(m,n,s)} + \beta_v)}{(\sum_{r=1}^{V} n_{(.),r}^{k,(y),-(m,n,s)} + \beta_r))} \end{aligned} \tag{1}$$

Here $Z_{-(m,n,s)}^s$ are the topic assignments for all tokens except the one corresponding to the n^{th} token in the m^{th} document in source s, $C_{-(m,n,s)}^s$ denotes the identity of the corresponding sources generating these token. $n_{m,(.)}^{k,(s),-(m,n,s)}$ is the frequency of the words assigned to the topic k in the document m of the source s when the token occurring in the n^{th} position in the m^{th} document in this source is removed. Similarly, $n_{(.),r}^{k,(y),-(m,n,s)}$ is the frequency of the r^{th} word in topic i in source y after removal of the n^{th} word in the m^{th} document in source s. Here, γ_y is the probability of the source being y.

This method depends on distribution of sources, for which we consider three different scenarios:

3.1 A-priori Source LDA

In this method, the distribution of sources from which the topics are drawn is fixed *a-priori*. This is especially useful in applications, where the relative importance of the sources is known *a-priori*. For example, for an event manager who wants to measure the crowd pulse regarding an event; the public opinion about that event might be more useful as compared to mere factual statistics about it. In such a scenario, social media websites might be given more importance than news websites.

3.2 Primary Source LDA

In this method, one of the sources is chosen as a primary source. To generate a token, if the word of the token is present in the vocabulary of the primary source, then the word distributions of the topics are drawn from the primary source. Otherwise, it is drawn from the source to which the token belongs. This is especially useful, when there is wide variance in the volume of data generated from different sources and the amount of data generated by some sources is very small. For instance, it is likely that Twitter will have far more tweets about an event than there are news articles from the New York Times. In such a scenario, the information obtained from other sources, can be used to denoise the data from an individual source and improve the learning of latent topics from that source.

3.3 Dirichlet Source LDA

In the event that there is no prior knowledge or information about the relative merits of sources, the distribution of sources can be drawn from a Dirichlet distribution. We refer to this as *Dirichlet Source LDA*. Assuming that the source distribution $\bar{\gamma}$ is generated from a Dirichlet distribution with parameter $\bar{\mu}$, the total probability of the model becomes:

$$P(W, Z, C, \theta, \phi, \bar{\gamma}; \bar{\alpha}, \bar{\beta}, \bar{\mu}) =$$
$$P(\bar{\gamma}; \bar{\mu}) \Pi_{s=1}^{S} \Pi_{i=1}^{K} P(\phi_i^s; \bar{\beta})$$
$$\Pi_{j=1}^{M^s} P(\theta_j^s; \bar{\alpha}) \Pi_{x=1}^{N^s} P(Z_{j,x}^s | \theta_j^s) P(C_{jx}^s | \bar{\gamma}) P(W_{j,x}^s | \phi_{{Z_{j,x}}^{C_{jx}^s}})$$

Using collapsed Gibbs sampling for inference, it can be shown that

$$\begin{aligned} &P(Z_{m,n}^s = k, C_{m,n,}^s = y | Z_{-(m,n,s)}^s, C_{-(m,n,s)}^s, W; \bar{\alpha}, \bar{\beta}, \bar{\mu}) \\ &\propto P(Z_{m,n}^s = k, Z_{-(m,n,s)}^s, C_{m,n}^s = y, C_{-(m,n,s)}^s, W; \bar{\alpha}, \bar{\beta}, \bar{\mu}) \\ &\propto \frac{(n_{m,(.)}^{k,(s),-(m,n,s)} + \alpha_k)}{(\sum_{i=1}^{K} n_{m,(.)}^{i,(s),-(m,n,s)} + \alpha_i)} . \frac{(q_y^{-(m,n,s)} + \mu_y)}{\sum_{x=1}^{S} (q_x^{-(m,n,s)} + \mu_x)} \\ &. \frac{(n_{(.),v}^{k,(y),-(m,n,s)} + \beta_v)}{(\sum_{r=1}^{V} n_{(.),r}^{k,(y),-(m,n,s)} + \beta_r))} \end{aligned} \tag{2}$$

In this equation, $q_y^{-(m,n,s)}$ is the number of tokens that are generated using word distributions of topics from source y, excluding the token in the n^{th} position of the m^{th} document in source s.

4. DATASETS

To evaluate these models, we collected feeds about the 2012 US Presidential Elections in the month of October 2012, from different sources comprising of news sources, social media, blogs, radio and discussion or social news forums. Some of these sources had special sections dedicated to the election. For others, we used specific query keywords like 'Election', 'USElection', 'USPresidentialElection', 'Obama', ‘Romney’, ‘Ryan’, ‘Biden’, ‘polls’, ‘campaign’ and ‘debate’ to extract election related feeds. The sources considered are:

1. Dedicated Blogs
   (a) The Election 2012 blog from the *Huffington Post*, which is American news website, content aggregator, and blog. [1].
   (b) *PoliticalWire*, a political blog based in the United States, published by Taegan Goddard. [2]
   (c) Daily political briefings on *Hotline.* [3]
   (d) Political blogs of the New York Times: *fivethirtyeight,* [4]*thecaucus,* [5]*krugman,* [6]*campaignstops,* [7] *takingnote.* [8]
2. Radio : We obtained election related stories from *NPR (National Public Radio)* using the queries enumerated above.
3. Social Media:
   (a) Twitter: We extracted tweets from the microblogging website Twitter containing at least one of the query words mentioned above
   (b) Reddit: We used the above query words to extract relevant posts and discussions about the elections
4. News
   (a) Political section of the *Huffington Post.* [9]
   (b) News from *Roll Call*, a newspaper published in Washington, D.C., United States, from Monday to Thursday when the United States Congress is in session and on Mondays only during recess[10]. We extracted daily news from *Political, Lobbying* and *Opinion* section from Oct 6, 2011.
   (c) Election related articles from the *New York Times.*[11]
   (d) Election related articles from the *Guardian.* [12]

We divide the data collected into two sets.

ElectionSet1. The first set comprises of mostly Twitter data from Oct 2, 2012 to Oct 9,2012 and data from other sources in the time period Oct 2, 2012 to Oct 22, 2012. For Twitter it comprises of more than 4 million tweets, more than 500 news stories from New York Times, more than 300 political blogs of New York Times and 100 blogs from Election 2012 blog of Huffington post, around 300 stories from NPR, around 650 news articles from Guardian and 400 articles from Political section of Huffington Post, around 47,000 discussion topics from Reddit, 65 news articles from Political Wire, 20 news from Roll Call, and around 200 articles from the political briefings on the Hotline Blog of National Journal.

ElectionSet2. The second batch comprises of feeds generated in all sources from Oct 20, 2012 to Oct 26,2012. For Twitter it comprises of more than 4.5 million tweets, more than 170 news stories from New York times, around 125 political blogs of New York Times and 60 blogs from Election 2012 blog of Huffington post, around 175 stories from NPR, around 250 news articles from Guardian and 165 articles from Political section of Huffington Post, more than 18,000 discussion topics from Reddit, 25 news articles from Political Wire, 80 news from Roll Call and around 60 articles from the political briefings on the Hotline Blog of National Journal.

5. EXPERIMENTS

In this section, we detail our experimental evaluation of the proposed methods. We will first present the metrics that we use to evaluate the proposed methods. Subsequently we will describe the experiments and results in detail.

5.1 Evaluation Metrics

The evaluation metrics used are:

Perplexity. Perplexity is the traditional metric for evaluating topic models. It is defined as:

$$perplexity = exp-\frac{\sum_{s=1}^{S}\sum_{d=1}^{M_s} log(p(w_d))}{\sum_{s=1}^{S}\sum_{d=1}^{M_s} N_d^s} \tag{3}$$

It measures how good a topic model is in generating the bag of words in the collection. A low value of perplexity is an indicator of good performance for the topic model that is being evaluated.

Topic Entropy. Perplexity finds the probability of generating a word using the model. Since our focus is on detecting topic models especially useful for applications related to document summarization, categorization of documents into topics and identification of correlated documents, we would prefer a scenario where one or a few topics can *maximally cover* or summarize a document. Hence we use another metric, *entropy* for evaluation. We define the topic entropy based evaluation metric as:

$$entropy(S) = exp-\frac{\sum_{s=1}^{S}\sum_{j=1}^{M_s}\frac{1}{N_j}\sum_{i=1}^{N_j} log(p(Z_{w_{j,i}}^s))}{\sum_{s=1}^{S} M_s} \tag{4}$$

Again, smaller the value of the entropy measure, better are the topics detected *i.e.*, smaller values of entropy indicate that the topics detected help to disambiguate or classify topics better.

5.2 Results and Evaluation

[1] http://www.huffingtonpost.com/news/election-2012-blog
[2] http://politicalwire.com/
[3] http://hotlineoncall.nationaljournal.com/
[4] http://fivethirtyeight.blogs.nytimes.com/
[5] http://thecaucus.blogs.nytimes.com/
[6] http://krugman.blogs.nytimes.com/
[7] http://campaignstops.blogs.nytimes.com/
[8] http://takingnote.blogs.nytimes.com/
[9] http://www.huffingtonpost.com/huff-wires/Politics.php
[10] http://www.rollcall.com/
[11] http://www.nytimes.com/
[12] http://www.guardiannews.com/

We evaluated all the proposed topic models using the two datasets for different values of number of topics K (5,7 and 10). Table 3 gives the numbering of sources (Source 0 corresponds to Twitter and so on).

For *a-priori LDA*, we consider two cases 1) the sources are picked from an arbitrarily chosen(random) probability distribution (*a-priori LDA1*) 2) when the probability of choosing each source is equal (*a-priori LDA2*). Dirichlet Source LDA is represented by *Dir LDA*. Each of these methods gives the word distribution of the global topic (*i.e.* the word distribution of a topic taking all sources under consideration). For each source, each of these methods also gives the word distribution for each local topic (i.e. word distribution of the topic unique to that source).

In this experiment, we evaluate all the *ProbLDA* topic models with the LDA baseline (applied for each source individually) for comparison.

To give an illustration of the results obtained, Table 2 shows the popular words belonging to the different global topics for the *Primary Source LDA* method with Twitter as the primary source for the *ElectionSet1* dataset at $K = 7$.

- Topic 1 deals with matters pertaining to jobs and performance in electoral colleges.
- Discussions on the tax plans and foreign policy, including the foreign policy of United States in the Middle East are included in Topic 2.
- The focus of Topic 3 seems to be millions of people watching the national debate and the candidates participating in it.
- Topic 4 includes feeds pertaining to GOP, truth and lies spoken during the election and Jon Stewart of the 'Daily Show' mocking Fox News' election night meltdown [13]
- Topic 5 features issues pertaining to economy, unemployment and women.
- Ohio where voting machines were tampered shows up in Topic 6, along with Michelle Obama and health related issues. It also includes feeds about NFL (National Football League), since NFL games were also being held during this period.
- Topic 7 includes mentions of media and Youtube (which provided live coverage of the election). It also includes mentions of the 'teaparty' and of 'Hugo Chavez' who won the presidential elections in Venezuela around the same time.

Table 3 gives the corresponding local topics in the different sources for Topic 5. We see that though the local topics are correlated to the global topics, they retain the characteristics of the individual sources.

T1	T2	T3	T4	T5	T6	T7
jobs	tax	poll	gop	class	real	media
bird	plan	america	lies	money	bird	youtube
america	days	support	report	economy	report	chavez
year	foreign	black	register	numbers	nfl	venezuela
policy	policy	candidate	support	unemployment	ohio	foreign
money	middle	foreign	money	bill	policy	twitter
college	east	watching	truth	women	post	care
performance	support	country	stewart	payne	michelle	bill
numbers	pbs	million	bush	attack	health	issues
million	hunger	live	fox	games	women	teaparty

Table 2: Some of the words (occurring with highest probability in the topics) characterizing the latent global topics detected using *Primary Source LDA* with Twitter as primary source on *ElectionSet1* at $K = 7$. Ti stands for topic i.

S.No	Source	Popular words				
0	Twitter	money	economy	job	unemployment	women
1	Nytimes	law	fiscal	growth	women	spending
2	NyBlogs	economy	abortion	florida	business	advertising
3	NPR	michelle	celeste	economy	unemployment	spending
4	HuffBLog	mandel	guinta	rape	child	abortion
5	Guardian	johnson	abortion	sexism	julia	healthcare
6	Reddit	binders	kitchen	women	hofstra	candy
7	Plo.Wire	florida	spending	economic	budget	candy
8	Roll Call	susan	spent	collins	budget	women
9	HuffWire	sexual	woman	virginia	spending	lopez
10	Hotline	desjarlais	advertising	mourdock	spending	guinta

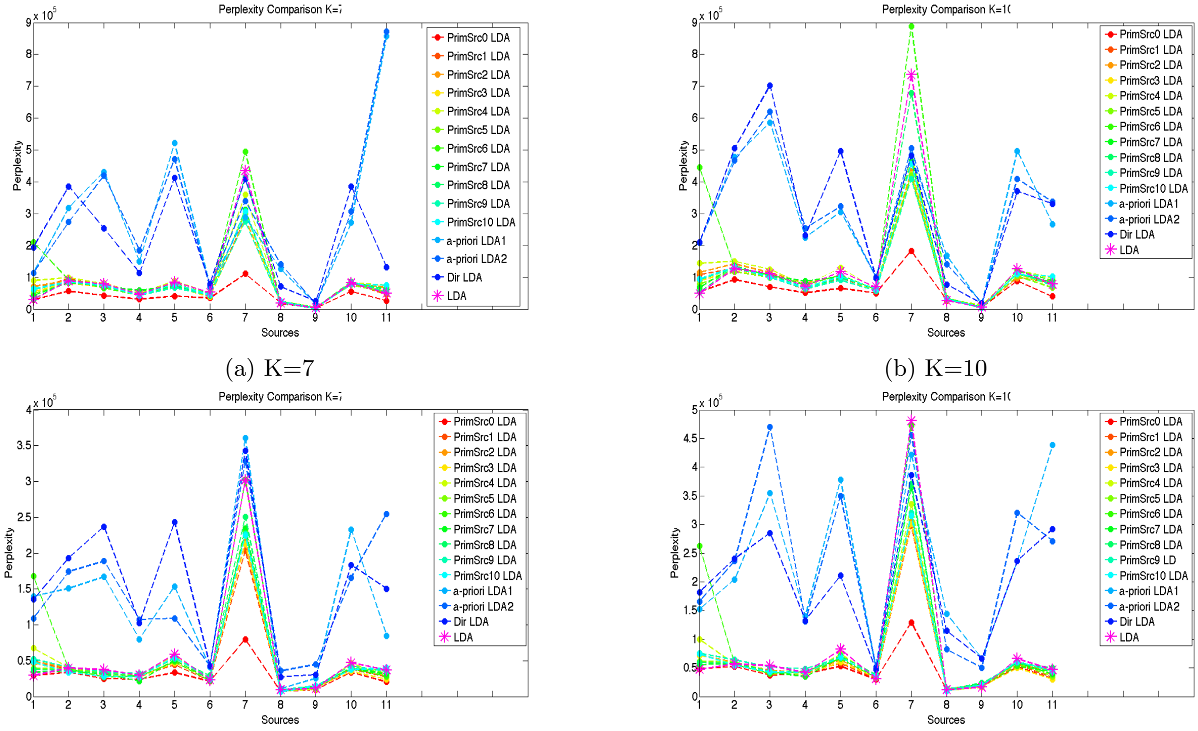
Table 3: Some of the words occurring with high probability in the local topics corresponding to global Topic 5 using*Primary Source LDA* with Twitter as primary source on *ElectionSet1* at $K = 7$.

Perplexity: Figure 2 gives the perplexity results for all the methods over the two datasets when the number of topics are 7 and 10 respectively. We observe that of the different *ProbLDA* methods, *Primary Source LDA method* outperforms *a-priori LDA* and *Dirichlet Source LDA* methods in all the cases studied. *Primary Source LDA method* with Twitter as the primary source gives the best perplexity results overall irrespective of the dataset or the number of topics and even outperforms LDA when applied to each source individually.

Entropy: Figure 3 gives the entropy results for the two datasets over all the methods. Similar to the results of using perplexity as the evaluation metric, we observe that amongst all the *ProbLDA* methods, Primary Source LDA with Twitter as the primary source appears to give the best result (lowest topic entropy) though its result is often comparable to running LDA on each source individually.

To summarize, among the *ProbLDA* methods, using Twitter as the primary source provides the best result (for both perplexity and topic entropy as evaluation metric) irrespective of the number of topics or the dataset studied. Its performance is equivalent or better than LDA applied on each source individually. Its superior performance can be attributed to the fact that it uses additional knowledge from other sources to learn the local topics of each individual source. This collective learning process helps in denoising the information mined from each individual source. Another distinct advantage of all the *ProbLDA* methods over LDA is that they give topic-to-topic correspondence between different sources and yet retain the distinct identity of each

[13] http://www.huffingtonpost.com/2012/11/08/jon-stewart-fox-news-election-meltdown-video_n_2092224.html



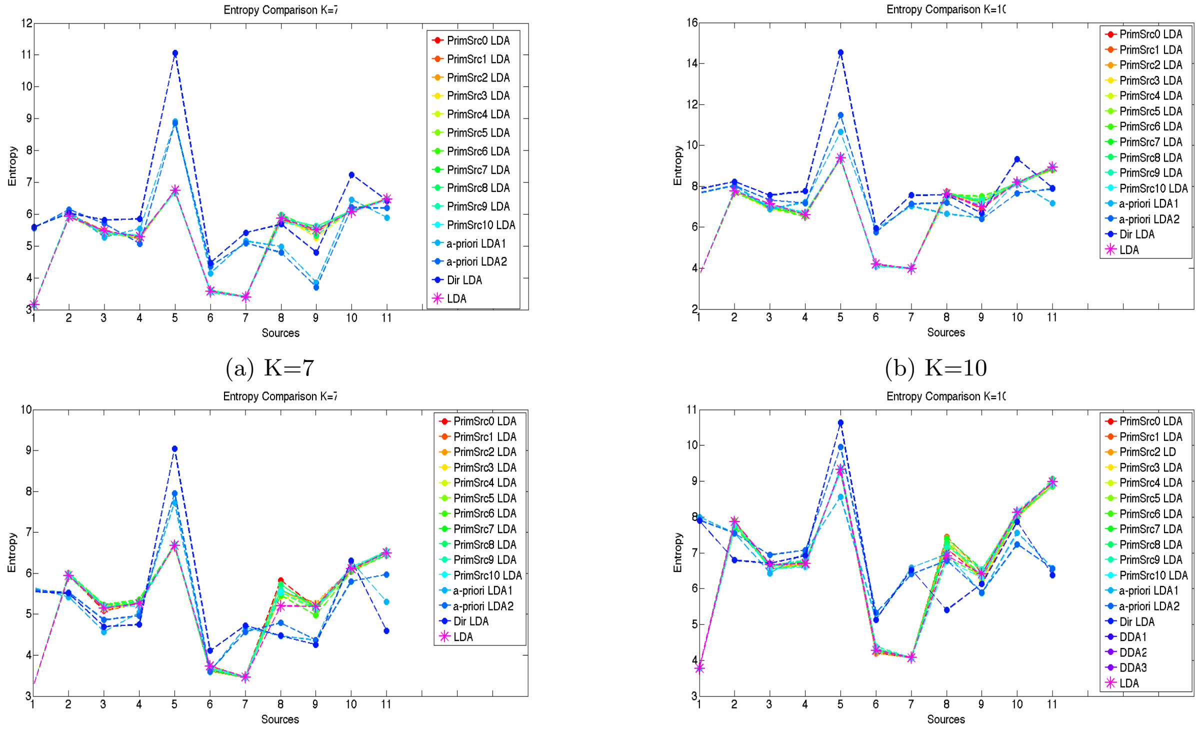
(a) K=7

(b) K=10

(c) K=7

(d) K=10

Figure 2: Comparison of perplexity for the Probabilistic Source LDA methods and LDA on each source individually for *ElectionSet1* and *ElectionSet2*. (a) and (b) give the perplexities for *ElectionSet1* at K=7, K=10 and (c) and (d) show the perplexities for *ElectionSet2* respectively.



(a) K=7

(b) K=10

(c) K=7

(d) K=10

Figure 3: Comparison of entropy for the Probabilistic Source LDA methods and LDA on each source individually for *ElectionSet1* and *ElectionSet2*. (a) and (b) give the topic entropies for *ElectionSet1* at K=7, K=10 and (c) and (d) show the topic entropies for *ElectionSet2* respectively.

individual source. This is especially useful in this age of information overload when it is becoming increasingly difficult to manually scan through different heterogenous sources, organize and correlate the information gathered.

6. DISCUSSION AND FUTURE WORK

In this paper we have addressed the challenging problem of mining useful information from different heterogeneous web sources, many of which are likely to be correlated. Therefore, apart from data compression, our objective is to find optimal latent topics leveraging the properties of each of the individual sources, and ensuring that there is topic-topic correspondence between the local topics of different sources. To the best of our knowledge, no prior work has tackled this problem. In this regard, we have proposed a novel topic modeling framework, *Probabilistic Source LDA* for this purpose. Probabilistic Source LDA by nature depends on the distribution of sources. For this we have defined three different methods – *a-priori Source LDA*, *Primary Source LDA* and *Dirichlet Source LDA*.

We have conducted extensive experiments using feeds extracted from a wide range of heterogeneous sources including social media, news sources, blogs and even radio transcripts, all pertaining to the US presidential elections in 2012. The proposed *Probabilistic Source LDA* method achieves topic-topic correspondence between different topics. Our results have shown that amongst the Probabilistic LDA methods, the Primary Source LDA with Twitter as the main source gives the best performance. We believe that the reason Twitter makes a good primary source is that it contains tweets with a wide coverage of topics, while news tends to be more focused on particular aspects. Thus, the secondary sources tend to help denoise the data from twitter, and thereby present meaningful topic models. We believe that the proposed method can have extensive use in applications like summarization of documents, identification of correlated documents in heterogeneous sources, topical categorization of documents and social search. As future work, we plan to generalize this model further, to take additional features like temporal dynamics into account and detect emerging topics.

7. REFERENCES

[1] H. Becker, M. Naaman, and L. Gravano. Selecting quality twitter content for events. In *ICWSM'11*, 2011.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[3] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[4] Hongbo Deng, Bo Zhao, and Jiawei Han. Collective topic modeling for heterogeneous networks. In *SIGIR '11*, pages 1109–1110. ACM, 2011.

[5] Aria Haghighi and Lucy Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 362–370, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[6] Thomas Hofmann. Probabilistic Latent Semantic Analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.

[7] Liangjie Hong, Byron Dom, Siva Gurumurthy, and Kostas Tsioutsiouliklis. A time-dependent topic model for multiple text streams. In *KDD '11*, pages 832–840. ACM, 2011.

[8] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification. *NIPS*, 2008.

[9] Timothy N. Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, pages 1–52, December 2011.

[10] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[11] Ivan Vulic, Wim De Smet, and Marie-Francine Moens. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval*, 16(3):331–368, 2013.

[12] Xiang Wang, Kai Zhang, Xiaoming Jin, and Dou Shen. Mining common topics from multiple asynchronous text streams. In *WSDM '09*, pages 192–201, New York, NY, USA, 2009. ACM.

[13] Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. Mining correlated bursty topic patterns from coordinated text streams. In *KDD '07*, pages 784–793, New York, NY, USA, 2007. ACM.

[14] ChengXiang Zhai, Atulya Velivelli, and Bei Yu. A cross-collection mixture model for comparative text mining. In *KDD '04*, pages 743–748. ACM, 2004.