DiscWord: Learning Discriminative Topics

Yu Jiang, Xian Li and Weiyi Meng
The Department of Computer Science
Binghamton University
Binghamton, NY USA
{yjiang5, xli2, meng}@binghamton.edu

***Abstract*—Topic modeling is a popular research topic and is widely used in text mining based applications. Many researchers realize that the learned topics in the LDA model, each as a multinomial distribution on the word vocabulary space, are often not intuitive in term of human recognition and communication. Based on our observation, given a topic, the most frequent words in it are usually less important than some words that are dedicated to it. In this paper, aiming at learning discriminative topics, we introduce a measure named *word discriminability* to capture a word's ability to identify different topics, and propose an iterative algorithm that is able to train and utilize word discriminability information during the topic learning process. Experimental results show that applying our method on the LDA topic model can improve its document classification accuracy significantly; the learned topics are more discriminative, and the top words of a topic are usually more representative.**

Keywords—discriminative topic, topic model, feature selection

I. INTRODUCTION

Topic model is a very important research topic in text mining. The generative topic model such as Latent Dirichlet Allocation (LDA) [1] is extremely popular in the text mining community. The primary advantage of generative topic model is that the coherence information between words, considered as the key to topic modeling, can be captured in the generative process. Besides, it can be easily applied to different kinds of applications, or extended by adding new latent random variables to support rich features like sentiment [2, 3], time [4, 5, 6] and geo-location identification, or integrating documents labels with topic model [7].

However, as the LDA topic model is applied to more and more applications, the limitation of LDA on its expressiveness has been noticed. Specifically, in high-level text analysis applications, people are interested in not only clustering texts into latent topics, but also discovering the exact meaning or contents of topics. To know the content of a topic is crucial to applications such as topic sentiment analysis [2] and topic topological study [8]. Therefore, topic representation with rich explicit content is desired. Some intuitive meanings to describe a topic explicitly include: a) use a group of theme keywords to represent a topic that is intuitive to human (e.g. twitter trending topics); and b) the selected theme words of topic should show the important features of the topic, not only probable, but also specific to the topic.

Let's use a real example to explain the idea. Table I shows an example LDA topic learned from a financial news corpus; its most probable words and a representative document snippet of this topic are shown. Two interesting findings are worth noticing: first, among the top ten most probable words in the 1st row (in descending order of the conditional probability of word on topic as shown in the 2nd row), there are noisy words like "said" and "late" that are common English words, and general words like "investors" and "markets" that are common in many documents of the corpus. These words are certainly not the representative words of the topic; second, obviously the document snippet is talking about different currencies, however, "currency" is only ranked at the seventh in the top ten probable word list and some important words like "yen" does not even appear in this list.

TABLE I. THE MOST PROBABLE WORDS OF AN LDA TOPIC AND A REPRESENTATIVE DOCUMENT SNIPPET ABOUT IT

Word	**dollar**	**euro**	said	debt	late	rating	**currency**	investors	markets
$P(w\|z)$	**0.037**	**0.021**	0.017	0.013	0.012	0.011	0.011	0.011	0.010
$P(z\|w)$	**0.49**	**0.42**	0.04	0.11	0.14	**0.39**	**0.44**	0.08	0.08

"*The U.S.* ***dollar*** *index climbed Monday as the* ***euro*** *gave back earlier gains against the* ***greenback****, with the two* ***currencies*** *advancing versus the Japanese* ***yen*** *as tensions between Japan and China over disputed islands showed signs of worsening.*"

This example shows that using the top probable words (which is widely used in many works [1,2,4,5,8, etc.]) to represent a topic is often not a good choice because they are not *discriminative*. The words that are common in many documents are often identified and ranked as representative features of topics due to their high frequency. Consequently, the actual representative features (i.e. theme words) of a topic that can differentiate this topic from the others are not included. As a result, it is hard for the users to obtain an explicit and correct idea about the topics covered in the corpus.

Continuing with the example in Table I, we have the third interesting finding. If we compute the conditional probability of topic on the words $P(z|w)$ (shown in the 3rd row), we can see that words "dollar", "euro", "rating" and "currency" have values that are significantly larger than the others, which means that the presence of these words has a much higher chance to indicate the presence of the topic. In other words, they are actually better candidate theme words of the topic than the other six words, since they are not only probable, but also deterministic to the topic. Topics learned with emphasis of such words are more likely to be discriminative from each other.

Now we define a topic as a "*discriminative topic*" if it satisfies the following two conditions:

- **Word sparsity**. The theme words of the topic can be easily distinguished from the other content words. This matches people's intuition that a topic is a group of correlated keywords. The topics learned using LDA have limited sparsity, i.e., it's difficult to select a small number of words to represent each topic clearly. This is because the topics learned by LDA, as a multinomial distribution on words, have no natural sparsity guarantee [9].
- **Topic discriminability**. The topic can be easily distinguished from the other topics. The topics learned in LDA are not discriminative enough due to an effect of the so-called "curse of dimensionality": with a fixed number of training samples (documents), the clustering quality decreases as the dimension of features (word vocabulary) increases. Especially, during the topic modeling process, the common words (e.g. "said" mentioned in Table I), due to their high frequencies in the documents, participate in many topics at the same time. Their presence a) increases the similarity between two different topics; b) reduces the significances of other words that are actually important to a topic. We call this effect as the "***dilution effect of words with high frequency***" in this paper.

Based on our observation, we believe that exploring the discriminative power of words' occurrences in predicating topics' occurrences in the topic modeling process would result in a significant improvement to the quality of the learned topics. It could leverage the impact of relevant words to a topic from the noises of other content words during the learning process, and help find the actual theme words of topics.

In this paper, aiming at learning discriminative topics and actual topic theme words, first we define the "*word discriminability*" (Section 2), which captures the ability of a word's occurrence in predicting the occurrence of a certain topic. Second, we propose a novel algorithm *DiscWord*, to learn and apply the word discriminability knowledge in the topic modeling process by dynamically updating the "weight" of words based on their discriminability values (Section 3). The algorithm is able to leverage the impact of the actual topic theme words during the learning process. Finally, we evaluate the performance of our proposed algorithm in a group of experiments on two corpora (Section 4). The contributions of this paper are:

- We propose the notion of "discriminative topic" that we believe is very important in the topic based high-level text mining. The most probable words of a topic are not necessarily the theme words. However, the words that are dedicated to a topic could be more important to the topic. In practice, distinguishing words of different importance is crucial in the process of topic modeling.
- We introduce a novel algorithm that iteratively learns the words' importance and apply them into the topic modeling process. As it favors those words with higher discriminative power in topics recognition, it is able to learn discriminative topics.
- Our experiments show that applying our proposed algorithm on top of LDA improves its accuracy for documents classification significantly, and the learned topics using our algorithm have higher quality, as they have more sparse word distributions, and are more discriminative from each other. Moreover, our method is able to find some topics that LDA fails to identify clearly.

II. PROBLEM ANALYSIS

In the Introduction, we explained our intuition that utilizing a word's ability to predict the presence of a topic can help us find more discriminative topics. In this section, we formally define the "word discriminability" as well as its evaluation metric to capture this intuition and we also explain the possible way to apply this idea to the topic modeling process.

In this paper, we use lower case letters w, d and z to denote word, document and topic random variables respectively, and capital letters W and D to denote the word vocabulary space and the observed document set, i.e. the corpus. A document is represented as a count vector of words in the vocabulary space $d = <n_{w,d}>$ for $w \in W$, where $n_{w,d}$ is the number of occurrences of w in document d. Suppose there are K topics in the corpus, each topic (z=k) can be characterized as a probability distribution (we also call it topic-distribution in this paper) on the vocabulary space W, i.e. $\boldsymbol{\varphi_k}=\{P(w|z=k) \mid w \in W\}$, where $P(w|z=k)$ is the conditional probability of word w on *the* k^{th} *topic*. The goal of topic modeling is to find the optimal parameter $\boldsymbol{\varphi} = \{\varphi_k\}$ that best fits the observed document set D.

A. Word Discriminability

In a generative topic model like LDA, the occurrence of every word is considered as drawn from a certain topic. We define the *Word Discriminability* as a word's ability to determine which topic it is drawn from. Note that the concept of "word discriminability" is independent of the document context of a word's occurrence. Evaluating the discriminability of a word w takes all of its occurrences in the corpus into consideration. From the definition above, we can see that evaluating word discriminability requires the prior knowledge of the topics. In this section, we first show how to compute the word discriminability with the assumption that topics are known and later in section 3 we will present an algorithm that evaluates the discriminability and learns the topics in an iterative manner.

To measure word discriminability with the assumption of the prior knowledge of the K topics, we use the Shannon entropy of the conditional probability distribution of the topic z on word w, as shown in Equation (1). Given the observation of w, it tells how much additional information (in terms of number of bits, i.e. $-ln(P(z=k|w))$) is needed to predict the topic assignment z of w. The more information that is needed, the more uncertainty of the prediction, and the less knowledge the word w has about the topic k. Thus the word discriminability (λ_w) of w is defined as the normalized inverse of the Shannon entropy as shown in Equation (2).

$$H_w = H(z|w) = -\sum_{k=1}^{K} P(z=k|w)\ln\left(P(z=k|w)\right) \quad (1)$$

$$\lambda_w = 1 - H_w/\ln(K) \quad (2)$$

The following two extreme cases show the reason why words' discriminability can be used for topic recognition.

(1) If a word w appears only when one topic is presented, say topic $z=j$, then we have $P(z=j|w) = 1$ and $P(z=k|w) = 0$ for all $k \neq j$. Therefore, the entropy is 0 and discriminability $\lambda_w = 1$; the occurrence of w indicates the occurrence of topic j with probability 1.

(2) If $P(z=k|w)$ is the same for all different topic z's, it has entropy value $ln(K)$ and discriminability $\lambda_w = 0$. The occurrence of w indicates nothing on which topic the word is drawn from.

B. Word discriminability integrated Topic model

We introduce the discriminability metric in order to leverage the importance of a word between its frequency and its dedication to a topic. In all existing topic models, each word occurrence has an equal contribution in the topic modeling process; so words with higher frequency have a larger impact. After utilizing word discriminability, a word with high frequency but participating in many topics should no longer be emphasized in the topic learning process. In contrast, a word with moderate frequency but with strong correlation with one specific topic should gain more attention. A straightforward example of applying word discriminability to topic learning can be described as follows. Considering topic modeling as a soft clustering problem on the corpus D, then each document d is a data point and each word in the vocabulary W is a feature, the value of feature w for d is $n_{w,d}$. Then the feature vector of d is $<n_{w,d}>$ for $w \in W$. Suppose for each word w, we obtain its discriminability λ_w and represent the discriminability of all words as a word discriminability vector $\lambda = <\lambda_w>$. Then the point-wise product of λ and d, i.e. $d' = \lambda \circ d$, produces a "weighted" feature vector for d. If we perform clustering on top of the weighted feature vectors of all documents in D, then the word discriminability knowledge is reflected in the clustering result (i.e. learned topics). The process is similar to performing a "weighted feature selection", i.e., assigning weight λ_w to each feature w, before starting the document clustering.

We can apply the word discriminability to a generative topic model in a similar way as in the above example. In general, the topic-modeling problem can be considered as an optimization problem that given the cost function $f(D, \varphi)$, try to find the optimal parameter φ (i.e. the probability distribution of topics) that minimizes the value of f, as in Equation (3). After adopting the word discriminability knowledge into the topic modeling process, the problem changes to finding the optimal $\hat{\varphi}$ for the weighted document set (see Equation (4)).

The perplexity [1] metric (see Equation (5)) is a good candidate for the cost function f: it measures how well the learned topic distributions (φ) predict the document set D. A low perplexity means that the learned topics generate the document set D with high probability. Here the N_D in the Equation (5) is the total number of word occurrences in the document set; it acts like a normalizing denominator. Note that other metrics can also used as the cost function f in Equation (4), as long as the weighted document set can be properly defined corresponding to the metric used.

$$\hat{\varphi} = argmin_{\varphi}\left(f(D,\varphi)\right) \quad (3)$$

$$\hat{\varphi} = argmin_{\varphi}\left(f(\lambda \circ D,\varphi)\right) \quad (4)$$

$$P(D;\varphi) = \exp\left(2, -\frac{1}{N_D}\sum_{d\in D}\log_2\left(P(d|\varphi)\right)\right) \quad (5)$$

III. An Iterative Topic Modeling Algorithm

In the previous section, we discussed the "word discriminability" and showed how it can be applied to the topic learning process. The discussion was based on the assumption of having prior knowledge about the topics. However, at the beginning of the learning process, we usually have no knowledge about what topics can be learned from the corpus. Therefore, a challenge is how to learn the topics and estimate the discriminability vector λ at the same time.

Our basic idea is that given the topic modeling results are already good, but still not good enough, why cannot we use the knowledge obtained in the existing result to find better result? Based on this idea, we add a "review" step to the iterative topic learning process. Initially, we have no knowledge of the topics as well as the words' discriminability; so we perform the iterative topic learning process on the corpus D in the same way as a regular generative topic model like LDA [1]. After a certain number of iterations, a "review" step is inserted to compute the discriminability of the words based on the topic distributions ($\boldsymbol{\varphi}$) learned so far. Though the estimated discriminability may not be accurate initially, it still provides information indicating if a word is discriminative on topics. Then start from the next iteration, we use the estimated discriminability values as the words' weights to generate the weighted document set $D' = \{d' = \lambda \circ d \mid d \in D\}$ and learn more discriminative topics based on it. This iterative process continues and the "review" step happens periodically until the learned topics become stable.

In this paper, we choose to implement our iterative algorithm on top of the collapsed Gibbs sampling algorithm [10] for the LDA topic model to evaluate its effectiveness. Nevertheless, our algorithm can be applied to various kinds of topic models with minor modification.

A. The collapsed Gibbs sampling algorithm for LDA

LDA [1] is a well-known generative topic model. In LDA, each topic φ_k is a multinomial distribution on the word vocabulary space, and each document d is drawn from a per document multinomial distribution θ_d on the topic space. LDA introduces two prior Dirichlet distributions β and α for the two multinomial distributions, respectively; so the learned topics will not be over-fitted by the training document set. We summarize the generative process of the LDA model as below:

1) Draw multinomial distribution φ_k over the vocabulary space W from Dirichlet distribution β for each topic $z=k$.

2) Draw multinomial distribution θ_d from Dirichlet distribution α for each document d.

3) For each position i in each document d,

   a) Draw a topic $z_{d,i}$ from θ_d.

b) Draw word $w_{d,i}$ from φ_k where $z_{d,i} = k$. (The word $w_{d,i}$ in document d at position i is considered as an observation of topic $z_{d,i}$).

The collapsed Gibbs sampling method [10] is widely used to learn the LDA topic model parameters θ and φ. The routine of the sampling algorithm strictly follows the above-mentioned generative process. In the algorithm, it uses the Monte Carlo method to estimate the posterior probability of random variable topic z conditioned on the observation of word w at position i_0 in document d_0, given the topic assignment of all the other words in the corpus. The posterior probability is computed based the Formula (8) below; it is the product of two parts: (1) the estimate of probability $P(w|z=k)$, as Equation (6), which is the probability of observing word w as an outcome of topic z; (2) the estimate of probability $P(z=k|d_0)$, as Equation (7), which is the probability of observing topic z in document d_0.

$$P_{-i}(w|z=k;\beta) = \frac{\sum_{d,i}\lambda_w I_{-i}(\mathrm{d,i,w,k})+\beta}{\sum_{v\in W}(\sum_{d,i}\lambda_w I_{-i}(\mathrm{d,i,v,k})+\beta)} \quad (6)$$

$$P_{-i}(z=k|d_0;\alpha) = \frac{\sum_{d=d_0,i}\lambda_{w_{d,i}} I_{-i}(\mathrm{d,i,\_,k})+\alpha}{\sum_{j=1}^{K}\sum_{d=d_0,i}\lambda_{w_{d,i}} I_{-i}(\mathrm{d,i,\_,j})+\alpha} \quad (7)$$

$$P(z=k|z_{-i},w;\alpha,\beta) \propto P_{-i}(w|z=k;\beta)P_{-i}(z=k|d_0;\alpha) \quad (8)$$

In Equations (6) and (7), the $I(d,i,w,k)$ is the Identity function, its value is 1 iff the word at position i in document d is w, and its topic assignment is k, otherwise it is 0; where the "_" in $I(d,i,\_,k)$ matches any word, the subscript $-i$ of I_{-i} indicates that the current position i_0 in document d_0 is excluded in the computation. The parameters α and β are the Dirichlet priors (usually some scalar values are used). The term $\lambda_{w_{d,i}}$ is the weight of the word at position i in document d, by default it is set to 1.0 [11] to all word occurrences in all documents. In our proposed algorithm it is replaced with the word's discriminability, since we are learning topics based on the "weighted" document set. The algorithm runs a sufficient number of iterations until convergence, and the topic model parameters (i.e. θ and φ) and other probability values can be computed based on the values of topic assignment $\boldsymbol{z}$ for all words occurrences in the corpus [11]. For example, the estimated conditional probability of topic k on word w, which is used in Equation (1) to compute word discriminability, can be computed using the following Equation (9):

$$P(\mathrm{z=k}|w) = \frac{\sum_{d,i}\lambda_w \mathrm{I(d,i,w,k)}+\beta}{\sum_{j-1}^{K}\sum_{d,i}\lambda_w \mathrm{I(d,i,w,j)}+\beta} \quad (9)$$

B. The DiscWord algorithm

Now we describe our proposed modified collapsed Gibbs sampling algorithm in detail. Since the word discriminability knowledge plays a key role in the algorithm, we name it as ***DiscWord*** for short.

The actual algorithm is quite neat (see Algorithm (1)). In the beginning, we randomly assign a topic to every word occurrence in the corpus (line 1); and also assign each word with equal discriminability value of 1.0. In the iteration from line 5-15, first, we adopt the Gibbs sampling step to estimate the topic assignment $\boldsymbol{z}$ of each word occurrence in the corpus (line 6-8). Second, we add a new "review" step that tries to update the word discriminability vector $\boldsymbol{\lambda}$ based on the learned topics so far (line 10-14). At last, we check whether the perplexity converges or not, if yes, the algorithm terminates; otherwise it continues the next iteration.

We now explain the "review" step in detail. The "review" step trying to update λ happens in every R iterations, where R is an empirical parameter. We first compute the new discriminability vector λ' based on the topics learned so far, and compute the updated perplexity value P' on the new weighted documents set $\lambda' \circ D$. We accept the change from λ to λ' if one of the two following conditions is satisfied (let ΔP denote $(P'-P)$ for simplicity):

1. $\Delta P < 0$, which means applying the new word discriminability vector would reduce the value of the perplexity. In such case, using λ' is better for learning discriminative topics; therefore we accept the new word discriminability vector λ'.
2. $\Delta P \geq 0$ *and* $rand() < exp(-1 * iter * \Delta P/P)$, i.e. we accept the update with certain probability even when the perplexity increases. The negative exponent is composed of two parts: $\Delta P/P$ is the perplexity change in percentage; *iter* is the current number of iterations. If the iteration number grows larger or the perplexity increases a lot, we are less likely to accept λ'.

The above mechanism of accepting new words' discriminability vector λ' adopts the idea of Simulated Annealing algorithm [12]; where $\Delta P/P$ and *iter* correspond to the state change and temperature change in simulated annealing, respectively. The mechanism has two advantages: the first condition prevents a lot of unnecessary changes on λ that are caused by sampling variance, especially in the last iterations of the learning process. The second condition provides a mechanism to avoid limiting the topic learning process at certain local minimum state of λ.

Finally, note that in the learned topics of our DiscWord algorithm, the probability $P(w|z)$ is based on the "weighted" document set, it does not reflect the word frequency in the original documents. However, this probability is able to reflect the importance of the word in the topic, combining both word frequency and discriminability factors. The top probable words of a topic generate very informative representation of the topic.

Algorithm 1. The DiscWord algorihtm for LDA

1	Sampling $<z_{d,i}>$ from the uniform distribution (1..K)
2	$\boldsymbol{\lambda} := <\boldsymbol{\lambda}_w>$, where $\boldsymbol{\lambda}_w := 1$ for each word w.
3	Compute perplexity P for $\lambda \circ D$ using Equation (5).
4	*iter* := 0
5	repeat
6	for each document *d* in D,
7	for *i* := 1 to N_d // N_d is the number of words in d.
8	draw $z_{d,i}$ from $P(z\|z_{-i})$ based on Formula (8).
9	update perplexity *P* for $\lambda \circ D$ using Equation (5).
10	if (++*iter* % R == 0) then
11	compute $\boldsymbol{\lambda}_w'$ for each w using Equation (2).
12	compute new perplexity *P'* for $\boldsymbol{\lambda}' \circ D$ using Equation (5).
13	if (*P'* < *P*) *or* (*rand*() < *exp*(-1 * *iter**(*P'* – *P*)/*P*)) *then*
14	$\boldsymbol{\lambda} := \boldsymbol{\lambda}'$ // accept the new discriminability vector
15	continue until perplexity P converges.

The DiscWord algorithm has some interesting properties compared to the original topic-learning algorithm for generative topic models:

- Word discriminability impacts the topic assignment of words occurrences; the effect of topic assignments is accumulated and later determines the words' *updated discriminability* values (line 11).
- A word with a higher discriminability has a larger impact on determining the topic assignment of other words (line 8), especially in the topics where the word has high probability. As a result, it attracts positively correlated words to join its topics gradually during the iterative process, forming a cluster finally.
- A word with a low discriminability is volatile to change its topic assignment (line 11). In the extreme case, a word with very low discriminability has its conditional probability on topic purely determined by the topic assignment of all of its positively correlated words. This gives the word the ability to change its discriminability "freely" as it is being attracted by its correlated words, until it finds its belonging topic (and gains a high discriminability then) or remains as an outlier to all topics (and remains at low discriminability). This feature guarantees that a word always gets the chance to increase its discriminability, i.e. our weighting algorithm only temporarily punishes, but never deletes a feature.

In practice, our algorithm effectively reduces the dilution effect of common words. Thus the learning process has a chance to explore the correlation between the actual topic theme words that were originally overlooked due to the large occurrences of common words.

Moreover, In the DiscWord algorithm, we enforce that the discriminability value of any word must be not less than a pre-defined small enough minimum value ε (e.g., 0.001), i.e. no word will be filtered out totally even its impact can be ignored. This setting is quite different from preprocessing techniques such as feature selection [13]. The reason is that a preprocessing mechanism could risk to filter out some common words that are actually important topic content. But in our algorithm, a word with low discriminability at the beginning is not eliminated and could later obtain a high discriminability after the words positively correlated with it form a topic.

IV. EXPERIMENTS

We compare our proposed algorithm DiscWord with the LDA model on two corpora, one is the well-known 20 newsgroups corpus (denotes as *20news*) [14], containing about 20,000 documents from 20 newsgroups; the other is a set of financial news articles in the year of 2011 crawled from the financial news website www.marketwatch.com (denoted as *finance*), containing about 2,500 unlabeled documents. We filter out stop words and words with total term frequency in the corpus less than a threshold, which is set as 10 for the newsgroups and 5 for the financial news corpus, and obtain word vocabularies of sizes about 15,000 and 6,000, respectively.

A. Document classification

Document classification is an important application of topic modeling. We compare the classification performance of LDA and DiscWord on the 20news corpus. The 20news corpus has pre-defined newsgroup labels for each document, we take them as the true classes of the documents. Our classification works as follows: first learn K topics from the corpus; then for each document, we choose the most probable topic as its topic label. For each topic label, we define its matching newsgroup label as the most frequent newsgroup label in the document collection with that topic label. A document is considered as correctly classified if its topic label matches its newsgroup label (true class). The classification accuracy result is shown in Table II. We vary K from 20 to 50, and find that in every case DiscWord outperforms LDA with a large margin (about 0.1 in absolute value on average), which is significant for a 20-class classification problem. This result implies that the topics learned by our DiscWord method match the human intuition (the newsgroup label) on topics much better. It is interesting to point out that as the number of topics increases, the classification accuracy of LDA tends to grow slowly. The reason is that as the number of topics grows, the occurrences of common words in a single topic decrease; thus the occurrences of the theme words comparatively "increase", making the topics learned by LDA more discriminative.

TABLE II. THE DOCUMENT CLASSIFICATION ACCURACY FOR THE 20NEWS CORPUS

#topic	20	25	30	35	40	45	50
DiscWord	**0.68**	0.64	0.65	0.65	0.66	0.65	0.65
LDA	0.55	0.53	0.54	0.55	0.56	0.57	**0.57**

B. Topic quality case study

In this section, we do a case study on real topics from the two corpora to compare the quality of the topics learned by LDA and DiscWord. In Table III, we list the most probable words and their probabilities of matching topics of the two models. Table III (a) shows the topic corresponding to the newsgroup label "*talk.politics.guns*". We can see that there are 5 common words "people", "government", "right", "state" and "control" in the top 10 words of the LDA topic; whereas in the DiscWord topic, these words are replaced with more topic specific words such as "firearm", "amendment" and "fbi". This suggests that the top words learned by the DiscWord method are more representative, revealing more features of the topic. Additionally, the probabilities of the top words of the DiscWord topic are much higher than those in the LDA topic, i.e. the DiscWord topic is more discriminative.

Table III (b) presents one low quality LDA topic and two related DiscWord topics that possibly match this LDA topic. We can see that the top words of the LDA topic all have similar probabilities, indicating the topic is not sparse and therefore not a good cluster (topic). Actually, its top words are composed of the top words of the two DiscWord topics on the right, one about "wal-mart" and the other about "libya" and "oil" - both of them have clear notions that match human's expectation of a topic. This observation indicates the LDA topic must be a composition of two topics. The DiscWord method is able to learn topics that LDA fails to identify.

TABLE III. THE DOCUMENT CLASSIFICATION ACCURACY FOR THE NEWSGROUPS CORPUS

(a) A pair of matching topics (newsgroup)

LDA		***DiscWord***	
gun	0.014	gun	0.032
law	0.007	guns	0.015
guns	0.007	crime	0.010
people	0.007	weapons	0.010
government	0.006	firearms	0.010
right	0.006	amendment	0.007
crime	0.005	fbi	0.007
weapons	0.005	militia	0.007
state	0.005	constitution	0.007
control	0.005	law	0.007

(b) A LDA topic and its two DiscWord matches (financial news)

LDA		***DiscWord***	
middle	0.016	-Match 1-	
east	0.016	mart	0.023
libya	0.012	wal	0.023
stores	0.011	stores	0.017
wal	0.011	community	0.012
mart	0.010	-Match 2-	
investors	0.010	oil	0.069
markets	0.010	east	0.047
closed	0.010	libya	0.047
gadhafi	0.009	middle	0.039

Next, we analyze the topic quality from another perspective, i.e. how does a word participate in multiple topics. We manually pick two words "oil" and "debt" from the finance corpus. These two words are very important topic words, and do not appear as frequently as corpus-common words like "marketwatch", "said" and "market". In Table IV, we list all the topics that contain either "oil" or "debt" as a representative word.

There are two LDA topics and three DiscWord topics related to the word "oil", which are matched in pairs in Table IV. The DiscWord topic #3 is about the oil-export country Libya that had a lot of violence in 2011. It does not have a matching topic in LDA. Even though #3 is a relatively small topic, all its representative words (except "oil") have high discriminability values; their impacts are leveraged by our DiscWord method. As a result, they are able to attract high-frequency word "oil" to participate in the topic. However, in the LDA model, the correlation of these words with "oil" is much smaller compared to other high-frequency words such as "energy", "prices" and "crude"; but the correlation between themselves are not strong enough to form a topic.

TABLE IV. TOPICS RELATED TO SELECTED WORDS

#	*LDA*	*DiscWord*
related topics of "oil"		
1	energy, oil, index, rose, stocks	energy, oil, gas, natural, arca
2	oil, prices, crude, barrels, said	oil, million, crude, barrels, energy
3	-no match found-	oil, east, libya, middle, unrest
related topics of "debt"		
4	debt, ceiling, said, obama, would	debt, ceiling, obama, house, trillion
5	european, debt, greece, euro, crisis	european, debt, greece, europe, crisis
6	**investors**, debt, since, global	*week, friday,* ***investors****, earnings*
7	**dollar**, **euro**, said, late, debt	***dollar, euro****, swiss, franc, currency*

"Debt" was a hot keyword in the financial news articles during the year of 2011 because of the "Greek government debt crisis[1]" and the "United States debt ceiling crisis[2]" at that time. There are four LDA topics and two DiscWord topics related to it, again matched in pairs in the bottom part of Table IV. The LDA topics #6 and #7 have no matching DiscWord topics, but they are actually of low quality: we can see that the LDA topic #6 is about "investors", "debt", and "global market", but the relationship between these keywords are not clear. The LDA topic #7 (which is the example we showed in Table I) talks about "dollar", "euro" and "debt"; we know that "debt" is certainly related to "euro" (the Greek crisis, see topic #5), and also probably related to "dollar" (the debt-ceiling crisis, see topic #4), but it is not clear why they are all put together to form one topic.

To investigate the reason, we find two DiscWord topics without the keyword "debt" that best match the LDA topics #6 and #7 (they are in italic font in Table IV), with the matching keywords highlighted in bold face. The DiscWord topic #6 is about "investors earnings", and the DiscWord topic #7 is about currencies. Clearly they are human intuitive topics. The biggest difference between them is that the LDA topics have common words like "since" and "said" which are not representative.

Last but not least, it is interesting to point out that the semantic relation between the representative words in the DiscWord topics is usually closer than that in the LDA topics, as is shown in the above examples. For example, in Table IV, "investors" and "debt" in LDA topic #6 have no direct relation, while "investors" and "earnings" in DiscWord topic #6 clearly have a strong relation. For another example, "oil" and "stocks" in LDA topic #1 have no direct relation, while "oil" and "gas" in DiscWord topic #1 have a strong relation, as they are different types of energy. This also supports the phenomena in the DiscWord algorithm that a discriminative word attracts positively correlated words to form topics. Even though LDA also utilizes word correlation information, the effect is not as good as DiscWord, mainly due to the dilution effect of the words with high frequencies.

C. Quantitative topic quality analysis

In this section, we first compare the *perplexity* (see Equation (5)) value of the two methods. DiscWord has a perplexity value of about **80%** of that of LDA (4513 vs. 5451) on the 20news corpus and a value of about **60%** of that of LDA (1698 vs. 2840) on the finance corpus. The result

[1] http://en.wikipedia.org/wiki/Greek_government-debt_crisis

[2] http://en.wikipedia.org/wiki/United_States_debt-ceiling_crisis_of_2011

indicates that the DiscWord method fits the texts in the corpora much better. In other word, adding the word discriminability knowledge into the learning process does improve topic-modeling quality.

Cluster Performance. Quantitatively, we evaluate topic-clustering result using two factors: inter-topic distance and in-cluster clustering quality. First, we compute the average KL-divergence of all topic pairs. KL-divergence is widely used to measure the distance between two probability distributions. Second, we evaluate the clustering quality of each topic using the topic sparsity measure defined in Equation (10) (similar to Equation (2)). The results are shown in Table V. We can see that the inter-topic distance of DiscWord topics is larger than that of LDA topics, and its topics are sparser (forming better clusters in the vocabulary space). We also compute the average of the total probabilities of the top ten words of each topic over all topics for each method. It is 0.09 for LDA and 0.14 for DiscWord. The top words in a DiscWord topic have much higher sum of probabilities than those in an LDA topic.

$$SP_k = 1 - H(w|z = k)/\ln(|W|) \quad (10)$$

TABLE V. TOPIC QUALITY COMPARISON RESULT

Metric	*LDA*		DiscWord	
	20-news	*Finance*	*20-news*	*Finance*
KL-divergence	3.69	3.63	3.76	3.69
SP	0.27	0.31	0.34	0.46

D. *Topic discriminability.*

To evaluate how discriminative the topics are, we select the top N words for each topic; then for each topic, we compute the maximum number of shared words with any other topic, which intuitively estimates the distance between a topic and its closest neighbor; at last, we compute the average distance for all topics, denoted as topics' "*closeness*", and use it to measure the hardness in distinguishing a random topic with its neighbors. By varying the parameter N, we obtain a chart of *closeness* on both corpora. We can see from Figure 1 that the DiscWord topics have much lower closeness values. On the 20news corpus, the closeness of DiscWord topics is as low as 2 for the top 20 words, that is three times smaller than that for the LDA topics.

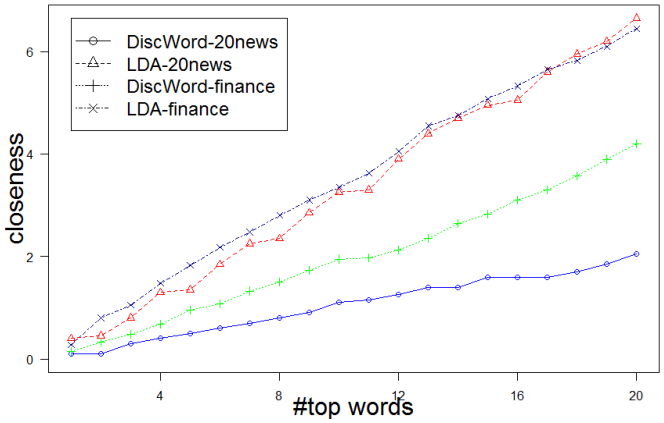

Fig 1. The number of shared top words between closest topic pairs

Note that the closeness of DiscWord topics on the finance corpus is higher than on the 20news corpus. This is because that the finance corpus is more domain specific and also has a smaller vocabulary, so the topics in this corpus share more representative words. In contrast, the closeness of LDA topics is insensitive to the number of topics, because a large portion of the top words of the LDA topics are corpus level common words that are invariant to the topic distribution in the corpus.

E. *Parameter tuning*

The parameter R in our iterative sampling algorithm controls the frequency of review steps. In implementation, we tried different values of R, ranging from 1 to 20; the learned topics are all similar with similar converging perplexity values. Given that topic modeling can be considered as a clustering problem on sparse high dimensional data, this result is expected and promising.

In terms of runtime efficiency, we would like to reduce the number of iterations in the topic learning process as well as the number of iterations of the review steps. But a small value of R would lead to a larger number of unnecessary iterations of the review step while a large value of R would incur more learning iterations based on some stale word discriminability vector. Based on our experiments, we find that R=5 provides a good balance for achieving good runtime efficiency.

V. RELATED WORK

Two lines of research are closely related to our work. One is applying word importance information into text mining and the other is topic modeling.

Word is the most obvious and most important source of information for text mining. Various kinds of metrics are defined based on word occurrence and frequency features for different kinds of tasks, including topic modeling [13]. Firstly, removing stop words is widely used in information retrieval as well as in topic modeling, which could be considered as a limited solution of filtering unrelated or less important words. Besides, the widely used word frequency based metrics such as document frequency (DF) and term frequency (TF) have been proved not very helpful in text clustering [15]. Also in [15], Liu et al propose to use term contribution (TC) based on document similarity as a metric to select important terms for document clustering. In our method, we measure the word discriminability based on its distribution over topics, rather than based on document similarities, which is fundamentally different. Another category of utilizing word importance information in topic modeling is feature selection. In [16], it assumes the topic information of documents is present and uses supervised dimensionality reduction technique (kernel PCA) to reduce the document set into lower dimensional space, and it then learns topics based on the reduced data. Our method differs from this method in that we do not assume any prior knowledge on the documents' belonging topics.

With respect to topic modeling, there are many research works that extend the basic LDA topic model [17]. Some incorporate metadata into the model, such as author [18], title, link [19], time [4,5,6], geographical location, and custom labels [7]. Some relax the assumption of LDA and add specific

constraints, such as topic correlation [20] or topic hierarchy [21]. And some put more emphasis on utilizing the information of document neighborhoods (i.e. local manifold structure on the high dimensional word vocabulary space) into topic modeling [21, 22, 23]. In this paper, we are not proposing a new topic model; instead we introduce an iterative algorithm that could be incorporated into most existing topic models (with small changes based on DiscWord) to learn more discriminative topics.

As far as learning discriminative topics is concerned, the research on topic sparsity is the most related to our work. In [3], to reduce the impact of noisy words that are not stop words, the authors introduce the notion of background topic that corresponds to the overall word distribution of the corpus, and learn the topics besides the background topic (i.e. each word occurrence has a fixed probability to be assigned to the background topic). In [9], the same background topic is introduced and the variational topics, which fall into the exponential probability distribution on the word vocabulary space, are then learned on top of the background topic. These models are able to learn topics with increased sparsity, and the impact of the non-topic-specific common words is reduced. However, because the background topic's probability distribution is strictly proportional to the corpus' word distribution, it also reduces the impact of the correlation between actual topic theme words in the modeling process. On the other hand, the impact of noise burst is amplified. In our work, we adopt a more flexible strategy to dynamically adjust the weight (i.e. discriminability) of each word based on the currently learned topics. As a result, our method is more intelligent at identifying the topic theme words and leveraging them gradually in the iterative learning process.

Finally, regarding topic quality evaluation, Wang and Blei [25] propose a "complexity" measure, i.e. the overall sum of the number of unique words for each topic in each document, to estimate the sparsity of topic model. In this paper, we introduce a simple metric called "*closeness*" to measure how close a topic connects to its neighbors on average. Both of these metrics offer some intuition on quantitatively measuring the topic modeling result from different perspectives. We think more such measures should be introduced as topic quality becomes more and more important in text mining applications.

VI. CONCLUSION

In this paper, we proposed an iterative topic learning method that can learn and utilize word discriminability information in the topic learning process. Our experimental results on two corpora show that our proposed method is able to learn discriminative topics that are more easily understandable by human, and their representations (i.e. the top words) provide more information about topics' content compared to existing techniques. In addition, our experimental results also show by incorporating our method into the LDA topic model, the accuracy of document classification can be improved significantly.

In the future, we plan to explore more applications of our method, such as developing quantitative measures on topic difference and connection based on the learned discriminative topics, in the context of topic evolution analysis. In addition, we would like to study in more depth the practical usage as well as the limitations of our proposed method.

VII. ACKNOWLEDGMENTS

This work was supported in part by NIH grant R01 LM010817 and NSF grant CNS-0958501.

REFERENCES

[1] D. M. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation". Journal of Machine Learning Research, 3:993–1022, January 2003.

[2] J. Yu, W. Meng, and C. Yu, "Topic sentiment change analysis". MLDM 2011, pp.443-457.

[3] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: modeling facets and opinions in weblogs". WWW 2007, pp.171–180.

[4] X. Wang, and A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends". SIGKDD 2006.

[5] L. AlSumait, D. Barbará and C. Domeniconi, "On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking". ICDM 2008.

[6] C. Wang, D. M. Blei and D. Heckerman, "Continuous Time Dynamic Topic Models". CoRR abs/1206.3298, 2012.

[7] D. Ramage, D. Hall, R. Nallapati and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora". ACL 2009, pp.248-256.

[8] Y. Jo, J. E. Hopcroft and C. Lagoze, "The web of topics: discovering the topology of topic evolution in a corpus". WWW 2011, pp.257-266.

[9] E. Eisenstein, A. Ahmed and E. P. Xing, "Sparse Additive Generative Models of Text". ICML 2011, pp.1041-1048.

[10] T. L. Griffiths, "Gibbs sampling in the generative model of Latent Dirichlet Allocation". 2002.

[11] T. L. Griffiths and M. Steyvers, "Finding scientific topics". Proceedings of the National Academy of Sciences of the USA, April 2004.

[12] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, "Optimization by Simulated Annealing." *Science* **220**, 671-680, 1983.

[13] Y. Yang, and O. J. Pedersen, "A Comparative Study on Feature Selection in Text Categorization". ICML 1997, pp.412-420.

[14] The 20 newsgroups dataset: http://qwone.com/~jason/20Newsgroups/

[15] S. Lacoste-Julien, F. Sha and M. I. Jordan, "DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification". NIPS 2008, 897-904.

[16] T. Liu, S. Liu, Z.Chen and W. Ma, "An Evaluation on Feature Selection for Text Clustering". ICML 2003, 488-495.

[17] D. M. Blei, "Probabilistic topic models". Commun. ACM (CACM) 55(4):77-84, 2012.

[18] M. Rosen-Zvi, T. L. Griffiths, M. Steyvers and P. Smith, "The author-topic model for authors and documents". Artificial Intelligence (2004), auai Press, 487–494.

[19] J. Chang and D. M. Blei, "Hierarchical relational models for document networks". Ann. Appl. Stat. 4(1), 2010.

[20] D. M. Blei and J. Lafferty, "A correlated topic model of science". Ann. Appl. Stat., 1, 1 (2007), 17–35.

[21] D. M. Blei, L. T. Griffiths, I. L. Jordan and B. J. Tenenbaum, "Hierarchical Topic Models and the Nested Chinese Restaurant Process". NIPS 2003.

[22] D. Cai, Q. Mei, J. Han, and C. Zhai, "Modeling hidden topics on document manifold". CIKM 2008, pp.911-920.

[23] S. Huh and S. E. Fienberg, "Discriminative Topic Modeling Based on Manifold Learning". TKDD 2012.

[24] H. Wu, J. Bu, C. Chen, J. Zhu, L. Zhang, H. Liu, C. Wang and D. Cai, "Locally discriminative topic modeling". Pattern Recognition 2012, 617-625.

[25] Wang, C. and Blei, D. M.: Decoupling Sparsity and Smoothness in the Discrete Hierarchical Dirichlet Process. NIPS 2009, pp.1982-1989.