Identifying Semantically Deviating Outlier Documents*

Honglei Zhuang[1], **Chi Wang**[2], **Fangbo Tao**[1], **Lance Kaplan**[3] and **Jiawei Han**[1]
[1]Department of Computer Science, University of Illinois at Urbana-Champaign
[2]Microsoft Research, Redmond [3]US Army Research Lab
{hzhuang3, ftao2, hanj}@illinois.edu,
wang.chi@microsoft.com, lance.m.kaplan.civ@mail.mil

Abstract

A document outlier is a document that substantially deviates in semantics from the majority ones in a corpus. Automatic identification of document outliers can be valuable in many applications, such as screening health records for medical mistakes. In this paper, we study the problem of mining semantically deviating document outliers in a given corpus. We develop a generative model to identify frequent and characteristic semantic regions in the word embedding space to represent the given corpus, and a robust outlierness measure which is resistant to noisy content in documents. Experiments conducted on two real-world textual data sets show that our method can achieve an up to 135% improvement over baselines in terms of recall at top-1% of the outlier ranking.

1 Introduction

The technology today has made it unprecedentedly easy to collect and store documents in an increasing number of domains. Automatic text analysis (*e.g.* document clustering, summarization, topic modeling) becomes more useful and demanded as the corpus size grows. Some trending domains (*e.g.* health records) call for a new analytical task, *mining outlier documents*: given a corpus, identify a small number of documents which substantially deviate from the semantic focuses of the given corpus. Outlier documents can provide valuable insights or imply potential errors. For example, an outlier health record from records of the same disease could indicate a new variation of the disease if it has an abnormal symptom description, or a medical error if it has an abnormal treatment description. A previous study (Hauskrecht et al., 2013) uses structured data in health records to show the importance of this application, and points out that further improvement should be achieved by leveraging text data.

Existing work has studied a related albeit different task, novel document detection (Kasiviswanathan et al., 2012, 2013; Zhang et al., 2002, 2004), where one aims to identify from a document stream if a newly arriving document is novel or redundant. In other words, this task assumes all the previous documents are known to be "normal", and only checks if a new document is novel. In our task, no document is known to be normal, and there could be multiple outliers in the corpus. Outlier detection (Chandola et al., 2009; Hodge and Austin, 2004) is a popular topic in data mining but few focus on text data. A study (Guthrie, 2008) identifies anomalous text segments in a document, but mainly based on writing styles. We focus on studying semantically deviating documents.

The problem of detecting outlier documents has its unique challenges. First, different words or phrases may be used to indicate the same semantic meaning, which introduces lexical sparsity. Second, finding proper words or phrases to characterize the corpus is non-trivial. Semantically frequent words or phrases can still be too general or too vague. Third, a document can carry extremely

*Research was sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1320617, IIS 16-18481, and NSF IIS 17-04532, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies of the U.S. Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2731–2740
Copenhagen, Denmark, September 7–11, 2017.

rich and noisy signals, most of which are not helpful to determine whether it is an outlier.

We tackle the problem of mining outlier documents in the following steps. We leverage word embedding (Mikolov et al., 2013) to capture the semantic proximities between words and/or phrases, in order to solve the sparsity issue. Then we propose a generative model to identify semantic regions in the embedded space frequently mentioned by documents in the corpus. The model represents each semantic region with a von Mises-Fisher distribution. We also learn a concentration parameter for each region with our model, and develop a selection method to identify semantically specific regions which can better represent the corpus, and filter regions with largely uninformative words.

As the final step, we design a robust outlierness measure emphasizing only the words or phrases in a document relatively close to the semantic focuses identified, and eliminating the noises and redundant information.

The remaining of the paper is organized as follows. Section 2 introduces the preprocessing of data sets and clarifies the notations. Section 3 proposes the methodology to mine outlier documents. Section 4 describes the experiment setup, Section 5 presents the results and Section 6 concludes the paper.

2 Preliminaries

In this section, we formalize the problem and then briefly describe the preprocessing step.

2.1 Notations

The notations used in this study are introduced here. A *document* is represented as a sequence $d_i = (w_{i1}, w_{i2}, \cdots, w_{in_i})$, where each $w_{ij} \in \mathcal{V}$ represents a word or phrase from a given vocabulary $\mathcal{V}$ and n_i denotes the length of the d_i. We refer to a set of documents as a *corpus*, represented as $D = \{d_i\}_{i=1}^{|D|}$.

Notice that w_{ij} may refer to a unigram word or a multi-gram phrase. Although it is nontrivial to appropriately segment a document into a mixed sequence of words and phrases, it is not the focus of our paper. A recently developed phrase mining technique (Liu et al., 2015) is used to extract quality phrases and segment the documents.

Word embedding provides vectorized representations of words and phrases to capture their semantic proximity. We assume there is an effective word embedding technique (*e.g.* (Mikolov et al., 2013)), $f : \mathcal{V} \mapsto \mathbb{R}^\nu$, where f is the transforming function that takes a word or a phrase as input and projects it into a ν-dimensional vector as its distributed representation. The semantic proximity between two words or phrases w and w' can be preserved by the cosine similarity between their embedded vectors:

$$CosSim\big(f(w), f(w')\big) = \frac{f(w) \cdot f(w')}{\|f(w)\| \times \|f(w')\|}$$

Problem definition. This work studies how to effectively rank documents in a corpus based on how much they deviate from the semantic focuses of the corpus. Given a set of documents D, our objective is to design an outlierness measure $\Omega : D \mapsto \mathbb{R}$, such that documents with larger outlierness $\Omega(d)$ semantically deviate more from the majority of D.

2.2 Preprocessing

We perform several steps of preprocessing to derive the input representation of each document in a given corpus.

Phrase mining. SegPhrase, a recently developed phrase-mining method (Liu et al., 2015), is utilized to automatically identify quality phrases in a corpus. After being trained in one corpus, SegPhrase is also capable of segmenting unseen documents into chunks of phrases with mixed lengths. We train SegPhrase on an external corpus D_e to obtain the list of quality phrases. Then for each corpus D given for outlier detection, we employ the trained SegPhrase to chunk each document into a sequence of words and quality phrases.

Word embedding. We adopt word embedding as a preprocessing step to capture the semantic proximity between words/phrases. Instead of using the raw text, similar to (Liu et al., 2015), we use the sequence derived from SegPhrase as input to the word embedding algorithm. In particular, *word2vec* (Mikolov et al., 2013) is utilized in our experiments, but can be seamlessly replaced by any other embedding results.

We run the embedding algorithm based on the external corpus D_e, the same corpus used in phrase mining. As D_e is sufficiently large, there are only few words or phrases in D which never appear in D_e, and are simply discarded in the experiments.

Stop words removal. We remove stop words, as well as the words or phrases ranked high within a certain quantile in terms of document frequency[1] (DF) in the external corpus D_e. Such words or phrases usually carry background noise, and obstruct outlier detection.

3 Mining Outlier Documents

Our framework consists of the following steps. First, we leverage a generative model to identify semantic "regions" in the word embedding space frequently mentioned by documents in the given corpus. Second, we develop a selection method to further remove semantics regions that are too general to properly characterize the given corpus, and only keep regions both frequent and semantically specific, denoted as "semantic focuses". Finally, we calculate the outlierness measure for each document based on the mined semantic focuses. We design a robust outlierness measure which is less sensitive to noisy words or phrases in documents.

3.1 Embedded von Mises-Fisher Allocation

We start with a generative model to identify the frequent semantic regions in the word embedding space.

Since we use cosine similarity to capture the semantic proximities between two words or phrases, the magnitude of the embedding vector of each word can be omitted in this part. We use $\mathbf{x}_{ij} = f(w_{ij})/\|f(w_{ij})\|$ to represent the unit vector with the same direction as the embedded vector of w_{ij}, and use $\mathbf{X}$ to represent the collection of all $\mathbf{x}_{ij}$ where $1 \leq i \leq |D|$ and $1 \leq j \leq n_i$.

In order to characterize a semantic region in the embedded space, we introduce von Mises-Fisher (vMF) distribution. The von Mises-Fisher (vMF) distribution is prevalently adopted in directional statistics, which studies the distribution of normalized vectors on a spherical space. The probability density function of the vMF distribution is explicitly instantiated by the cosine similarity. It is an ideal distribution for our task because we use cosine similarity to measure the semantic proximity. Moreover, as we will see later, it empowers us to characterize how specific each semantic region is, which is helpful in further identification of semantic focuses for outlier detection.

[1]Document frequency of a word (or phrase) is defined as number of documents where this word or phrase appears.

We first introduce the formalization of the von Mises-Fisher distribution.

Von Mises-Fisher (vMF) distribution. A ν-dimensional unit random vector $\mathbf{x}$ (*i.e.* $\mathbf{x} \in \mathbb{R}^\nu$ and $\|\mathbf{x}\| = 1$) follows a von Mises-Fisher distribution $\text{vMF}(\cdot|\boldsymbol{\mu}, \kappa)$ if the probability density function follows:

$$p(\mathbf{x}) = C_\nu(\kappa) \exp\left(\kappa \boldsymbol{\mu}^\top \mathbf{x}\right)$$

where $C_\nu(\kappa) = \kappa^{\nu/2-1} / (2\pi)^{\nu/2} I_{\nu/2-1}(\kappa)$; and $I_{\nu/2-1}(\cdot)$ is the modified Bessel function of the first kind; $(\nu/2 - 1)$ is the order.

The two parameters in the vMF distribution are the mean direction $\boldsymbol{\mu}$ and the concentration parameter κ respectively, where $\boldsymbol{\mu} \in \mathbb{R}^\nu, \|\boldsymbol{\mu}\| = 1$ and $\kappa > 0$. The distribution concentrated around the mean direction $\boldsymbol{\mu}$, and is more concentrated if the concentration parameter κ is larger.

Embedded von Mises-Fisher allocation. We propose a generative model by regarding each document as a bag of normalized embedded vectors, analogous to the bag-of-word representation of documents utilized in typical topic model (*e.g.*, LDA (Blei et al., 2001)). The major difference is that the data to be generated is now a bag-of-normalized-embedded-vectors for each document, and should be generated from a mixed vMF distribution instead of a mixed multinomial distribution.

A formalized description of the model is summarized as follows:

$$\begin{aligned}
\boldsymbol{\mu}_t &\sim \text{vMF}(\cdot|\boldsymbol{\mu}_0, C_0), && t = 1, 2, \cdots, T \\
\kappa_t &\sim \text{logNormal}(\cdot|m_0, \sigma_0^2), && t = 1, 2, \cdots, T \\
\boldsymbol{\pi}_i &\sim \text{Dirichlet}(\cdot|\boldsymbol{\alpha}), && i = 1, 2, \cdots, |D| \\
z_{ij} &\sim \text{Categorical}(\cdot|\pi_i), && j = 1, 2, \cdots, |d_i| \\
\mathbf{x}_{ij} &\sim \text{vMF}(\cdot|\boldsymbol{\mu}_{z_{ij}}, \kappa_{z_{ij}}), && j = 1, 2, \cdots, |d_i|
\end{aligned}$$

where $T > 0$ is an integer indicating the number of semantic regions, namely the number of vMF distributions in our mixture model.

We regularize the vMF parameters by the following prior distributions. We assume the mean direction $\boldsymbol{\mu}_t$ of each vMF distribution is generated from a prior vMF distribution $\text{vMF}(\cdot|\boldsymbol{\mu}_0, C_0)$, while the concentration parameter κ_t is generated from a log-normal prior $\text{logNormal}(\cdot|m_0, \sigma_0^2)$. A similar design is also adopted in (Gopal and Yang, 2014).

Parameter inference. We infer the parameters by Gibbs sampling. Because both the von Mises-Fisher distribution and the Dirichlet distribution

have conjugate priors, we can integrate out parameters $\boldsymbol{\mu}_t$ and $\boldsymbol{\pi}_i$ and develop a collapsed Gibbs sampler of z_{ij}:

$$P(z_{ij}=t|\mathbf{Z}^{-ij},\mathbf{X},\boldsymbol{\kappa};\boldsymbol{\alpha},m_0,\sigma_0^2,\boldsymbol{\mu}_0,C_0)$$
$$\propto\frac{(n_{it}^{-ij}+1+\boldsymbol{\alpha}^{(t)})C_\nu(\kappa_t)C_\nu\big(\|C_0\boldsymbol{\mu}_0+\kappa_t\mathbf{x}_{\cdot t}^{-ij}\|\big)}{C_\nu\Big(\Big\|C_0\boldsymbol{\mu}_0+\kappa_t\big(\mathbf{x}_{\cdot t}^{-ij}+\mathbf{x}_{ij}\big)\Big\|\Big)}$$

where $n_{it}^{-ij}=\sum_{j'}^{|d_i|}\delta(z_{ij'}=t)-\delta(z_{ij}=t)$ is the number of words in the i-th document being assigned to the t-th von Mises-Fisher distribution without taking w_{ij} into account; $\mathbf{x}_{\cdot t}^{-ij}=\sum_{i'}^{|D|}\sum_{j'}^{|d_i|}\mathbf{x}_{i'j'}\delta(z_{i'j'}=t)-\delta(z_{ij}=t)$ is the sum of word vectors assigned to semantic region t without counting w_{ij}. Here $\delta(\cdot)$ is the indicator function.

We can also derive a collapsed Gibbs sampler for concentration parameters κ_t's:

$$P(\kappa_t|\mathbf{Z},\mathbf{X},\boldsymbol{\kappa}^{-t};\boldsymbol{\alpha},m_0,\sigma_0^2,\boldsymbol{\mu}_0,C_0)$$
$$\propto\frac{C_\nu^{n_{\cdot t}}(\kappa_t)}{C_\nu\big(\|C_0\boldsymbol{\mu}_0+\kappa_t\mathbf{x}_{\cdot t}\|\big)}\text{logNormal}(\kappa_t|m_0,\sigma_0^2)$$

where $n_{\cdot t}$ is the number of words in semantic region t.

While sampling z_{ij} is relatively trivial, sampling κ_t is not straightforward. Similar difficulty is also mentioned in (Gopal and Yang, 2014). We employ a Metropolis-Hasting algorithm with another log-normal distribution centered at the current κ_t value as the proposal distribution.

After obtaining a sample from the posterior distribution of z_{ij}'s and κ_t's, we can easily obtain the MAP estimate of mean directions $\boldsymbol{\mu}_t$'s and the mixing distribution of each documents $\boldsymbol{\pi}_i$:

$$\hat{\boldsymbol{\mu}}_t=\frac{C_0\boldsymbol{\mu}_0+\kappa_t\mathbf{x}_{\cdot t}}{\|C_0\boldsymbol{\mu}_0+\kappa_t\mathbf{x}_{\cdot t}\|},\quad \hat{\boldsymbol{\pi}}_i=\frac{n_{it}+\boldsymbol{\alpha}^{(t)}}{n_{i\cdot}+\sum_t\boldsymbol{\alpha}^{(t)}}$$

Discussions. We notice that there are some topic models (Das et al., 2015; Batmanghelich et al., 2016) proposed for similar data, where words are represented as embedding vectors. Our model is proposed independently for the purpose of identifying semantic focuses, which serves the task of outlier detection. Existing models may lack signals for the following outlier detection steps and hence cannot be directly plugged in. However, it is possible to adapt certain models to the outlier detection task.

3.2 Identifying Semantic Focuses

The semantic regions learned from the Embedded vMF Allocation model provide a set of candidates frequently mentioned by documents in the corpus. However, not all of them are semantic focuses of the corpus — some are too general to distinguish outlier and normal document.

We notice that uninformative semantic regions (*e.g.* a semantic region containing {"percent", "average", "compare", ...}) tend to have more scattered distribution over embedded vectors, possibly because of the diverse context of their usage. In contrast, corpus-specific semantic regions are more concentrated, (*e.g.* a semantic region containing {"drugs", "antidepressant", "prescription", ...}). Modeling semantic regions by vMF distributions provides us with a parsimonious signal to characterize how concentrated a semantic region is, *i.e.* the concentration parameter κ_t. This allows us to simply filter unqualified semantic regions with too small concentration parameters and obtain high-quality semantic focuses. Let a binary variable ϕ_t ($t=1,2,\cdots,T$) indicate whether the t-th vMF distribution is a semantic focus. Suppose a user specifies a threshold parameter $0\leq\beta\leq 1$. We can determine ϕ_t by estimating the log-normal distribution that generates all κ_t's, $logNormal(\hat{m},\hat{\sigma}^2)$, where

$$\hat{m}=\frac{1}{T}\sum_t\log(\kappa_t),\ \hat{\sigma}^2=\frac{1}{T}\sum_t\Big(\log(\kappa_t)-\hat{m}\Big)^2$$

Set $\hat{F}_\kappa(\cdot)$ to be its cumulative distribution function. We assign $\phi_t=1$ for semantic regions with $\kappa_t\geq F_\kappa^{-1}(\beta)$, and filter all the other semantic regions as $\phi_t=0$.

Although parameter β needs to be set manually, our experiments suggest the performance is not quite sensitive to its value.

3.3 Document Outlierness

In this subsection, we start with a straightforward definition of outlierness based on the mined semantic focuses. Then we present several refinements to improve its robustness.

Baseline outlierness measure. A straightforward intuition is to assume outlier documents averagely have fewer words or phrases drawn from semantic focuses. To estimate this, we first need to calculate the probability of each word being drawn

from the semantic focuses.

$$P\big(\phi_{z_{ij}} = 1|\mathbf{x}_{ij}, \boldsymbol{\pi}_i\big) = \frac{\sum_t \phi_t \boldsymbol{\pi}_i^{(t)} \text{vMF}(\mathbf{x}_{ij}|\boldsymbol{\mu}_t, \kappa_t)}{\sum_t \boldsymbol{\pi}_i^{(t)} \text{vMF}(\mathbf{x}_{ij}|\boldsymbol{\mu}_t, \kappa_t)}$$

It is then possible to estimate the expected percentage of words *not* drawn from semantic focuses in each document as the outlierness:

$$\Omega_{\text{sf}}(d_i) = 1 - \frac{1}{|d_i|}\sum_{j=1}^{|d_i|} P(\phi_{z_{ij}} = 1|\mathbf{x}_{ij}, \boldsymbol{\pi}_i) \quad (1)$$

However, due to the noisiness in text data, this assumption oversimplifies the characterization of outlier documents. In practice, we observe the following two issues: lexically general words/phrases, and noisy content in documents.

Penalizing lexically general words and phrases. Not all words or phrases close to semantic focuses are strong indicators of normal documents. General words (*e.g.* "science") can happen to be semantically close to a semantic focus, but are not as specific as most other words close to it (*e.g.* "medical research"). Therefore, we utilize a background corpus D_{bg} to calculate the specificity of the word. Assuming the actual mention of the word can be chosen from either the general background, or a corpus-specific vocabulary, we write down the probability that a word is corpus-specific to be:

$$P\big(\lambda_{ij}|w_{ij}\big) = \frac{nd(w_{ij})/|D|}{nd(w_{ij})/|D| + nd_{bg}(w_{ij})/|D_{bg}|}$$

where $nd(w) = |\{d_i|w \in d_i, d_i \in D\}|$ is the number of documents in D containing word w; $nd_{bg}(w) = |\{d_i|w \in d_i, d_i \in D_{bg}\}|$ is the number of documents containing word w in the background corpus D_{bg}; λ_{ij} is a binary random variable indicating whether w_{ij} is specific enough.

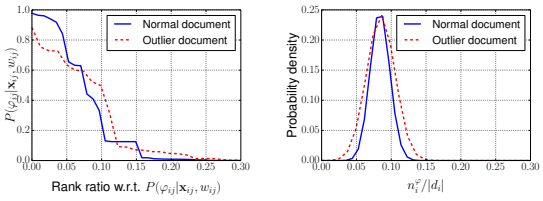
For each word, we define the word is *orthodox* if the word is not only semantically close to a semantic focus of the corpus, but also sufficiently specific. We then define the probability that a word or phrase w_{ij} in document d_i is orthodox as:

$$P(\varphi_{ij}|\mathbf{x}_{ij}, \boldsymbol{\pi}_i w_{ij}) = P\big(\phi_{z_{ij}}|\mathbf{x}_{ij}, \boldsymbol{\pi}_i\big) P\big(\lambda_{ij}|w_{ij}\big)$$

where $\varphi_{ij} = 1$ indicates that w_{ij} (or equivalently $\mathbf{x}_{ij}$) is orthodox.

Now, we can define a second outlierness measure as the expected percentage of words that are *not* orthodox.

$$\Omega_{\text{e}}(d_i) = 1 - \frac{1}{|d_i|}\sum_{j=1}^{|d_i|} P(\varphi_{ij}|\mathbf{x}_{ij}, \boldsymbol{\pi}_i, w_{ij}) \quad (2)$$



(a) Ranked orthodox probability $P(\varphi_{ij}|\mathbf{x}_{ij}, \boldsymbol{\pi}_i, w_{ij})$ (b) Probability distribution of random variable $n_i^{\varphi}/|d_i|$

Figure 1: Comparison of a normal document and an outlier document in a news corpus ("Health" topic).

Noisy content in documents. We present the second issue of normal documents with an example. We compare a normal document in a corpus of New York Times news articles with tag "Health", to another document originally from another corpus, but with its outlierness calculated with regard to the semantic focuses of the "Health" corpus.

In Figure 1(a), we show the distribution of inferred orthodox probability $P(\varphi_{ij} = 1|\mathbf{x}_{ij}, w_{ij})$ by ranking the words or phrases according to their probability value. We can observe that the outlier document barely has any words or phrases surely orthodox, while the normal document has 5% of words or phrases with a probability no less than 0.8 to be orthodox. However, if we simply take the average, these two documents become indistinguishable as the average is substantially dominated by the "tail" where most words or phrases in either documents are clearly not orthodox. Let n_i^{φ} be a random variable indicating the true number of orthodox words or phrases in document d_i. Since n_i^{φ} follows a Poisson-Binomial distribution, we can plot the probability distribution of n_i^{φ} normalized by the length of the document, as shown in Figure 1(b). It can be observed that the difference between the normalized expectation $\mathbb{E}[n_i^{\varphi}]/d_i$ of two documents is insignificant. Therefore, the measure described in Equation (2) will be unable to tell the difference between these two documents.

This example illustrates why the strategy of taking the average over the whole document can make mistakes, and also provides an important insight. As long as a document has a (potentially small) portion of words or phrases that are highly certain to be orthodox, it should not be considered as an outlier. Based on the above observation, we propose a third outlierness measure.

Orthodox quantile outlierness. We define a *quantile-based outlierness* definition to rank document outliers. Notice that the distribution of random variable n_i^{φ} follows a Poisson-Binomial distribution, which is the total number of success trials when one tosses a coin for each word or phrase in the document to determine whether it is orthodox with probability $P(\varphi_{ij}|\mathbf{x}_{ij}, w_{ij})$.

Moreover, we define the first $\frac{1}{1-\theta}$-quantile of the Poisson-Binomial distribution of n_i^{φ} as:

$$q_\theta(n_i^{\varphi}) = \sup_q \{q : P(n_i^{\varphi} \geq q) \geq \theta\} \quad (3)$$

where $0 < \theta < 1$ is a given parameter close to 1. Intuitively, it measures the maximum lower bound of n_i^{φ} we can guarantee with confidence θ.

Based on Equation (3), we can give a formalized definition of our proposed outlierness:

$$\Omega_{\theta\text{-q}}(d_i) = 1 - \frac{q_\theta(n_i^{\varphi}) + 1}{|d_i| + 1} \quad (4)$$

where the $\frac{1}{1-\theta}$-quantile is normalized by the document length with a smoothing constant. The cumulative probability distribution of a Poisson-Binomial distribution can be efficiently calculated by dynamic programming (Chen and Liu, 1997).

The advantage of the last proposed outlierness measure is that it emphasizes more on the highly orthodox words or phrases and eliminates the noise from a number of relatively uncertain ones.

4 Experiment Setup

4.1 Data Sets

New York Times News (NYT). We collected 41,959 news article published in 2013 from The New York Times API[2]. Each article is assigned with a unique label indicating in which section the article is published, such as Arts, Travel, Sports, and Health. There are totally 9 section labels in our collected data set. We treat papers in each section as a corpus D. Thereby we have a set of corpora $\mathcal{D} = \{D_s\}$, without overlapping documents. We also have an external news data set D_e crawled from Google news, with 51,114 news article published in 2015 without any label information.

ArnetMiner Paper Abstracts (ARNET). We employ abstracts of papers published in the field of computer science up to 2013, collected by ArnetMiner (Tang et al., 2008), and assign each paper into a field, according to Wikipedia[3]. We use papers from a set of domains to serve as an external corpus D_e, while papers in other domains form different corpora $\mathcal{D} = \{D_s\}$. Each domain (*e.g.*, data mining, computational biology, and computer graphics) forms a corpus D_s respectively. Again, notice that the corpora do not have overlapping documents with each other.

A summary is presented in Table 1.

Table 1: Data set statistics.

Data set	Corpus D		External corpus D_e	
	Avg. $\lvert D\rvert$	Avg. $\lvert d\rvert$	$\lvert D_e\rvert$	Avg. $\lvert d\rvert$
NYT	4,662.11	592.66	52,114	471.63
ARNET	2,930.60	137.21	11,463	152.17

Benchmark generation. Since we do not have true labels for outliers in a corpus, we use injection method to generate outlier detection benchmark. For each data set, we randomly select a corpus $D_s \in \mathcal{D}$ and mark all of its document as "normal documents". We then randomly select another corpus $D'_s \in \mathcal{D}, D'_s \neq D_s$, to inject ω documents from D'_s into D_s and mark them as outliers. We confine ω to be a small integer less than 1% of the size of $|D_s|$. More concretely, ω is an integer uniformly sampled from $(0, 0.01|D_s|]$.

For each data set, we randomly generate 10 outlier detection benchmarks, and evaluate the overall performance by the average performance on all the benchmarks.

4.2 Methods Evaluated

We compare the performances of the following methods.

Cosine similarity based. We characterize each document as a vector, and use the negative average cosine similarity between each document and the corpus as outlierness. We use two different ways to vectorize documents: TF-IDF weighted, and paragraph2vec (Le and Mikolov, 2014). The two methods are denoted as TFIDF-COS and P2V-COS respectively.

KL divergence based. We represent each document as a probability distribution, and the entire corpus as another probability distribution. Then we use the KL-divergence between each document and the entire corpus as the outlierness. We also

[2] http://developer.nytimes.com/docs

[3] https://en.wikipedia.org/wiki/List_of_computer_science_conferences

use two different ways to calculate the probability distribution. The first is to estimate the unigram distribution for each document and the entire corpus respectively, denoted as UNI-KL. The other is to first perform LDA on the entire corpus with 10 topics, and then infer topical allocation distribution of each document and the entire corpus. This method is represented as TM-KL.

Our method Our quantile based method is denoted as VMF-Q. We also provide two baselines derived from our own method as an ablation analysis. One method abandons the quantile based outlierness but use the expected orthodox percentage as Equation (2), denoted as VMF-E. The other method further removes the penalty on lexical general words and phrases, using Equation (1), denoted as VMF-SF.

4.3 Evaluation Measures

In most outlier detection applications, people are more concerned with recall. We measure the performance by *recall at a certain percentage*. More specifically, we compute the recall of outlier detection if the user checks a certain percentage r of the top-ranked documents in the output results. Since in our benchmark generation, the percentage of outliers does not exceed 1%. Therefore, the perfect results for any $r \geq 1\%$ should be 1.0.

We choose r to be 1%, 2%, and 5% respectively and evaluate different methods with recall at top-r (percentage). We also report the performance in terms of mean average precision (MAP).

4.4 Parameter Configurations

All benchmark data sets are preprocessed as described in Section 2. In the NYT data set we remove words or phrases within top 20% with respect to document frequency, while in the ARNET data set we remove the top 10%. The document frequency is calculated based on a background corpus D_{bg}, which is the same as the external corpus of NYT. Word embedding are trained on the external data set D_e using code of Mikolov *et al.* (Mikolov et al., 2013) with default parameter configurations, where the embedded vector length is set to 200. For paragraph2vec, we learn the length-100 vectors for each document along with the external data set to guarantee sufficient training data.

For the prior vMF distribution, we set $C_0 = 0.1$, a sufficiently small number so the prior distribution is close to a uniform distribution. $\boldsymbol{\mu}_0$ is set as a normalized all-1 vector. We also set $m_0 = log(100)$, and $\sigma^2 = 0.01$. The total number for Gibbs sampling is set to be 50 times of the total count of z_{ij}'s (*i.e.* $\eta = 50$). The number of vMF distributions T is set to 20 in the NYT data set and 10 in the ARNET data set respectively, due to the smaller sizes of corpora in the ARNET data set.

To determine semantic focuses, we set threshold parameter $\beta = 0.55$ for both data sets. The confidence parameter θ in outlierness calculation is set to 0.95 in both data sets. Our experiments later will show the performance is relatively robust to different configurations of both parameters.

Table 2: Performance comparison of different outlier document detection methods. All results are shown as percents.

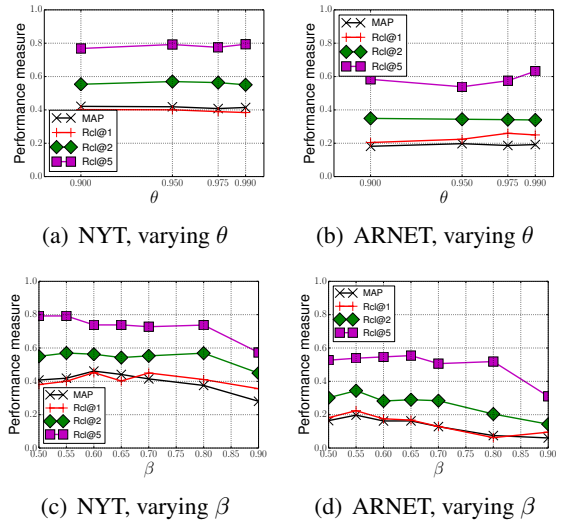
Data set	Method	MAP	Rcl@1%	Rcl@2%	Rcl@5%
NYT	TFIDF-COS	05.03	04.73	06.72	14.72
	P2V-COS	22.07	23.45	44.64	66.18
	UNI-KL	10.28	11.92	16.32	31.34
	TM-KL	14.51	16.50	16.50	24.67
	VMF-SF	33.70	31.03	44.45	62.60
	VMF-E	36.57	35.91	49.41	67.56
	VMF-Q	**41.88**	**56.99**	**63.29**	**79.23**
ARNET	TFIDF-COS	08.99	15.40	18.75	30.23
	P2V-COS	07.39	10.51	14.78	24.14
	UNI-KL	07.46	14.13	22.26	39.40
	TM-KL	10.09	12.04	15.37	20.24
	VMF-SF	10.69	12.05	22.58	44.51
	VMF-E	10.51	12.67	25.92	45.37
	VMF-Q	**19.74**	**22.40**	**34.40**	**53.87**

5 Results

We present the experimental results in this section.

Performance comparison. Table 2 shows performance of different outlier document detection methods. It can be observed that our method outperforms all the baselines in both data sets. In both data sets, VMF-Q can achieve a 45% to 135% increase from baselines in terms of recall by examining the top 1% outliers. Generally, performances of most methods are lower in the ARNET data set comparing to NYT, potentially because the relatively short document lengths and more technical terminologies in ARNET.

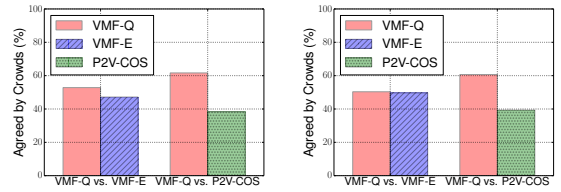
Ablation analysis. Both refinements of the outlierness measure benefits the performance. Specifically, by changing the average based outlierness to quantile based outlierness, the recall@1% can be improved by 50-75%, and the recall@5% can also be improved by more than 17%.



(a) NYT, varying θ (b) ARNET, varying θ

(c) NYT, varying β (d) ARNET, varying β

Figure 2: Performance of outlier document detection with different parameter configurations.



(a) Health (b) Art

Figure 3: Crowd evaluation to compare different outlier detection methods on two corpora in NYT data set.

Sensitivity studies of parameters. We study if our proposed method is sensitive to the confidence parameter θ and filtering threshold parameter β. We compare the performance of VMF-Q by varying each parameter on both data sets. Figure 2(a) and 2(b) show that the performance is not very sensitive to different values of θ, as long as θ is sufficiently large (close to 1). Figure 2(c) and 2(d) show that the performance is relatively stable when β is between 0.5 and 0.7, but drops a little when β is set to larger value.

Human judgments. We compare VMF-Q to VMF-E and P2V-COS respectively by crowdsourcing, *without* artificially inserting "outliers". We conduct this experiments on two corpora in NYT data sets with topic "Health" and "Art" respectively. To compare two methods, we randomly select pairs of documents d_i and d_j such that both are ranked as top-10% outliers by at least one method, but their orders in the two rankings disagree. We conduct the experiments on CrowdFlower. Online crowd workers are given d_i and d_j as well as other documents in the corpus, and are asked to judge which one of d_i and d_j deviates more from the corpus. For each corpus, we select 200 pairs of documents.

Before taking the questions, each crowd worker needs to go through at least 10 "test questions" which we know the correct answer. These questions are constructed by taking one document from the corpus as d_i and another document not from the corpus as d_j. Therefore, the one not from the corpus should be the answer. A crowd worker needs to achieve no less than 80% of accuracy to be eligible to work on actual questions, and the accuracy needs to be maintained over 80% during the work, which is measured by "test questions" hidden in actual questions. Each question is answered by 3 workers. The final answer is determined by majority voting.

Figure 3 presents the results. On both corpora, there are significantly more workers tend to agree with VMF-Q comparing to P2V-COS, with significance level $\alpha = 0.05$. This further verifies that our method VMF-Q can achieve better performance than the P2V-COS baseline. On the other hand, on both data sets we can still observe more workers favoring VMF-Q than VMF-E, but the difference is not as large as the difference between VMF-Q and P2V-COS.

Case study. We also conduct a case study to show how our proposed method outperforms other baselines. Table 3 shows two pairs of documents in "Health" corpus of NYT data set. The left two columns show some comparing methods and their higher ranked outlier documents. The row of "Crowds" shows the outlier document chosen by human workers from the crowdsourcing platform, with a consensus of opinions from multiple workers.

In the first document pair, document A is about gun control policy and is substantially irrelevant to "Health" topic, while document B is about lung infection cases. Document A is a significant outlier, and VMF-Q and VMF-E also agree with our intuition. However, paragraph2vec (P2V) ranks document B higher, probably because it tries to summarize the entire document.

In the second document pair, document B is clearly *not* an outlier as the story is about a new book of AIDS. In comparison, document A dis-

Table 3: Case study of documents in "Health" corpus of NYT data set. We present several pairs of documents and how different methods rank the pair. The "Outlier" column indicates the document ranked higher in the outlier document ranking generated by the corresponding methods, and the row "Crowds" shows the ranking given by human evaluators.

Method	Outlier	Document A	Document B
P2V-COS	Doc B	*CHICAGO (AP) States with the most gun control laws have the fewest gun-related deaths, according to a study that suggests sheer quantity of measures might make a difference ...*	*A prominent Scottish bagpiping school has warned pipers around to world to clean their instruments regularly after one of its longtime members nearly died of a lung infection ...*
VMF-E	Doc A		
VMF-Q	Doc A		
Crowds	Doc A		
P2V-COS	Doc B	*ATLANTA There's more evidence that U.S. births may be leveling off after years of decline. The number of babies born last year only slipped a little, ...*	*Young men in a state prison for juveniles and professors of library science from the University of South Carolina have joined forces to fight AIDS with a graphic novel ...*
VMF-E	Doc B		
VMF-Q	Doc A		
Crowds	Doc A		

cussing U.S. population is an outlier. However, a great part of document B is about the content of the book, which confuses baselines P2V and VMF-E, as both methods tend to summarize the entire document and highly relevant words like "AIDS" are overwhelmed by the majority of the document. The only method that agrees with human annotators is VMF-Q.

6 Conclusion

In this paper, we propose a novel task of detecting document outliers from a given corpus. We propose a generative model to identify semantic focuses of a corpus, each represented as a vMF distribution in the embedded space. We also design a document outlierness measure. We experimentally verify the effectiveness of our methods. We hope this work provides insights for further studies on outlier document texts in specific domains, and in more challenging settings such as detecting outliers from crowdsourced data.

References

Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Samuel Gershman. 2016. Nonparametric spherical topic modeling with word embeddings. In *ACL*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2001. Latent dirichlet allocation. In *NIPS*, pages 601–608.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15:1–15:58.

Sean X Chen and Jun S Liu. 1997. Statistical applications of the poisson-binomial and conditional bernoulli distributions. *Statistica Sinica*, pages 875–892.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *ACL*, pages 795–804.

Siddharth Gopal and Yiming Yang. 2014. Von mises-fisher clustering models. In *ICML*, pages 154–162.

David Guthrie. 2008. *Unsupervised Detection of Anomalous Text*. Ph.D. thesis, University of Sheffield.

Milos Hauskrecht, Iyad Batal, Michal Valko, Shyam Visweswaran, Gregory F Cooper, and Gilles Clermont. 2013. Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics*, 46(1):47–55.

Victoria J Hodge and Jim Austin. 2004. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126.

S. P. Kasiviswanathan, G. Cong, P. Melville, and R. D. Lawrence. 2013. Novel document detection for massive data streams using distributed dictionary learning. *IBM J. Res. Dev.*, 57(3-4):1:9–1:9.

Shiva P Kasiviswanathan, Huahua Wang, Arindam Banerjee, and Prem Melville. 2012. Online l1-dictionary learning with application to novel document detection. In *NIPS*, pages 2258–2266.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196.

Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. 2015. Mining quality phrases from massive text corpora. In *SIGMOD*, pages 1729–1744. ACM.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *KDD*, pages 990–998. ACM.

Jian Zhang, Zoubin Ghahramani, and Yiming Yang. 2004. A probabilistic model for online document clustering with application to novelty detection. In *NIPS*, pages 1617–1624.

Yi Zhang, Jamie Callan, and Thomas Minka. 2002. Novelty and redundancy detection in adaptive filtering. In *SIGIR*, pages 81–88. ACM.