Faceted Models of Blog Feeds

Lifeng Jia
Department of Computer Science
University of Illinois at Chicago
851 S. Morgan Street
Chicago, IL, USA, 60607
ljia2@uic.edu

Clement Yu
Department of Computer Science
University of Illinois at Chicago
851 S. Morgan Street
Chicago, IL, USA, 60607
cyu@uic.edu

Weiyi Meng
Department of Computer Science
Binghamton University
Binghamton, NY 13902
meng@cs.binghamton.edu

ABSTRACT

Faceted blog distillation aims at retrieving the blogs that are not only relevant to a query but also exhibit an interested facet. In this paper we consider personal and official facets. Personal blogs depict various topics related to the personal experiences of bloggers while official blogs deliver contents with bloggers' commercial influences. We observe that some terms, such as nouns, usually describe the topics of posts in blogs while other terms, such as pronouns and adverbs, normally reflect the facets of posts. Thus we present a model that estimates the probabilistic distributions of topics and those of facets in posts. It leverages a classifier to separate facet terms from topical terms in the posterior inference. We also observe that the posts from a blog are likely to exhibit the same facet. So we propose another model that constrains the posts from a blog to have the same facet distributions in its generative process. Experimental results using the TREC 2009-2010 queries over the TREC Blogs08 collection show the effectiveness of both models. Our results outperform the best known results for personal and official distillation.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - retrieval models, selection process;

General Terms

Algorithms, Experimentation, Languages

Keywords

Faceted Blog Distillation, Personal and Official Facets

1. INTRODUCTION

To explore information seeking behaviors in blogosphere, TREC 2009 [9] introduced faceted blog distillation that aims at, for a given query q in a blog search context, retrieving the blogs that are relevant to q and exhibit a given facet. In this paper, we are interested in a pair of facets: personal vs. official. Personal blogs are normally written by bloggers to describe various topics related to their personal experiences while official blogs are increasingly written by companies for PR or marketing purposes. A blog (i.e. an RSS feed) is a set of blog posts. We use the term *feed* to represent a blog and the term *post* to represent a blog post in a feed.

Existing personal and official blog distillation works are dictionary based [2, 7, 8], heuristic based [4, 11] and classification based [5, 6, 13, 15, 17]. In contrast, we employ topic modeling techniques. We observe that 1) the topics in posts are normally expressed by some terms (called *topical terms*), such as nouns, while the facets of posts are normally revealed by other terms (called *facet terms*), such as pronouns and adverbs; and 2) the facet terms frequently used in personal posts are different from those in official posts. Let us illustrate these two observations with an example.

Example 1. Given a query "*parenting*", one excerpt d_1 = "***I*** *am a big fan of babywearing and so* ***I*** *was elated when* ***I*** *received a Sleepy Wrap to try.* ***I*** *thought it was so cute that it came in a little pouch bag (for storage).* ***Anyway****, what exactly is a Sleepy Wrap?*" and another excerpt d_2 = "*A coalition of concerned individuals and organizations has started NEPI, the National Effective Parenting Initiative. Its goal is to have every child in our country* ***effectively*** *and* ***humanely*** *raised by parents who receive the best possible parenting education, training and support.*", d_1 is from a personal post while d_2 comes from an official post. The terms in bold obviously do not show any topics but facets. Specifically, the pronoun "*I*" and the interjection "*Anyway*" show the personal facet while the adverbs "*effectively*" and "*humanely*" show the official facet.

Motivated by our observations, we propose two models that calculate the probabilistic distributions of topics and those of the two facets within posts. Moreover, all the existing works indicated above treated posts independently and ignored the facet correlation among all the posts from a feed. We observe that a post from a feed f is likely to exhibit the same facet as other posts from f. It is intuitively reasonable, because all the posts from a feed are usually written by a blogger at different times. To leverage such a correlation, one of the two models constrains all the posts from a feed to have the same facet distribution in its generative process.

This paper has three contributions. 1) Our work is the first study that employs generative models to solve personal and official blog distillation. 2) Our work is the first study that leverages the facet correlation of the posts from a feed into the calculation of their personal and official facets. 3) Experimental results show our models are robust and effective and our best results outperform the best known results for personal and official blog distillation.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505657.

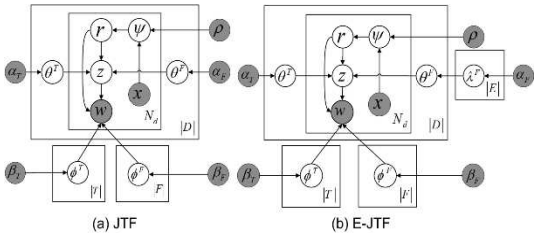

Figure 1: The plate notations of JTF and E-JTF.

2. RELATED WORK

Various approaches have been proposed to study personal and official blog distillation. The methods in [2, 6, 7, 17] made the assumption that personal (official) posts are likely to have opinionated (factual) contents. They indirectly measured the extents of posts exhibiting the personal (official) facet by measuring those of posts being opinionated (factual). Specifically, the extents of posts being opinionated or factual are calculated either by the opinionated and factual lexicons built by mutual information metric [2, 7] or by various opinionated and factual classifiers [6, 17]. Li et al. [8] built a personal and official lexicon using information gain measure and then used this lexicon to measure the personal and official facets of posts. Heuristics built on parameters, such as the number of occurrences of first person pronouns, are used to calculate the personal and official facets of posts in [4, 11]. Various classifiers are built in [5, 13, 15] to directly categorize posts into personal or official ones. Our work differs from the above existing works in two aspects: 1) unlike the methods in those works, we use the topic modeling techniques to compute the facets of posts; and 2) their works treated posts independently while one of our models constrains the posts from a feed to exhibit the same facet.

3. JOINT TOPIC FACET MODEL

We now present a joint topic facet model (called JTF) that computes the probabilistic distributions of topics and those of facets within posts. Specifically, given a set of feeds with respect to (w.r.t.) a query where each feed is a set of posts, a corpus is formed by pooling the posts from the feeds. In the JTF model, the terms (unigrams) of a post in the corpus are assumed to be generated from a mixture of some latent topics and two latent facets: personal and official.

3.1 Generative Process of the JTF Model

We first describe JTF's generative process. Let us define some notations for ease of presentation. Let E denote a set of feeds, $E = \{e_1, \ldots, e_{|E|}\}$ where each feed e_i is a set of posts. A corpus of posts is formed by pooling the posts from E. Let D denote the corpus, $D = \{d_1, \ldots, d_{|D|}\}$ where $d_i \in D$ is a post from a feed $e_{d_i} \in E$. Let V be the term vocabulary of D.[1] Each post d has N_d terms where $w_{d,n}$ denotes the n^{th} term in d. Let W be the set of observed terms in D, $W = \{w_{d,n} | d \in D; n = 1 \ldots N_d\}$. The terms in W act as either topical terms or facet terms. Let $\Psi = \{\psi_{d,n} | d \in D; n = 1 \ldots N_d\}$ be the set of the probabilities for all the terms in W acting as facet terms. Specifically, $\psi_{d,n}$ is the probability of $w_{d,n}$ acting as a facet term. Let $R = \{r_{d,n} | d \in D; n = 1 \ldots N_d\}$ be the set of the binary variables, each of which indicates whether a term in W acts as a topical term or a facet term. Specifically, $w_{d,n}$ acts as a facet (topical) term if $r_{d,n} = 1$ ($r_{d,n} = 0$). Let $Z = \{z_{d,n} | d \in D; n = 1 \ldots N_d\}$ be the set of the topics or the facets assigned to the terms in W. Specifically, $z_{d,n}$ is the facet (topic) assignment to $w_{d,n}$ if $r_{d,n} = 1$ ($r_{d,n} = 0$). Assume that there are a set of latent topics T and two facets F in D. $\Theta^F = \{\theta^F_{d,f}\}_{|D|\times|F|}$ is a matrix where $\theta^F_{d,f}$ is the probability of a post d exhibiting the facet $f \in F$ and the row vector θ^F_d is the probabilistic distribution of d exhibiting all the facets F. $\Theta^T = \{\theta^T_{d,t}\}_{|D|\times|T|}$ is the matrix where $\theta^T_{d,t}$ is the probability of d exhibiting the topic $t \in T$ and the row vector θ^T_d is the probabilistic distribution of d exhibiting all the topics T. $\Phi^F = \{\phi^F_{f,v}\}_{|F|\times|V|}$ is a matrix where the row vector ϕ^F_f is the probabilistic distribution over all the terms in V for the facet f and $\phi^F_{f,v}$ is the probability of the term v for f. $\Phi^T = \{\phi^T_{t,v}\}_{|T|\times|V|}$ is a matrix where the row vector ϕ^T_t is the probabilistic distribution over all the terms in V for the topic t and $\phi^T_{t,v}$ is the probability of the term v for t. Now we present the generative process of the JTF model (see the plate notation in Figure 1(a)).

---

1. For each facet $f \in F$, draw $\phi^F_f \sim Dirichlet(\beta_F)$;
2. For each topic $t \in T$, draw $\phi^T_t \sim Dirichlet(\beta_T)$;
3. For each post $d \in D$,
   1) Draw $\theta^F_d \sim Dirichlet(\alpha_F)$;
   2) Draw $\theta^T_d \sim Dirichlet(\alpha_T)$;
   3) For each term in d, say $w_{d,n}$,
      i) Set $\psi_{d,n} \leftarrow g(.)$;
      ii) Draw $r_{d,n} \sim Bernoulli(\psi_{d.n})$;
      iii) if $r_{d,n} = 0$ // *$w_{d,n}$ acts as a topical term.*
         a) Draw $z_{d,n} \sim Multinomial(\theta^T_d)$;
         b) Draw $w_{d,n} \sim Multinomial(\phi^T_{z_{d,n}})$;
      iv) else // *$w_{d,n}$ acts as a facet term.*
         a) Draw $z_{d,n} \sim Multinomial(\theta^F_d)$;
         b) Draw $w_{d,n} \sim Multinomial(\phi^F_{z_{d,n}})$;

In the generative process, $\psi_{d,n}$ is the probability that controls whether the term $w_{d,n}$ in d is generated from a facet or a topic and it is set by a function $g(\cdot)$ (see Step 3.3.i). Theoretically, we should use Beta priors as $g(\cdot)$ in the (unsupervised) generative process. The studies [12, 16] proposed two models that compute the topics and the sentiments of posts. In their models, a term is generated from either a topic or a sentiment. So their models set the probability of a term in a post acting as a topic term. Their studies showed that using classifiers in their posterior inferences to calculate these probabilities yields better performance than deriving these probabilities from the adopted Beta priors in their generative processes, because fully unsupervised topic models are unable to separate sentiment terms from topical terms well [16]. In this work, we want to separate topical terms from facet terms and we believe that it is hard to differentiate them in an unsupervised manner either. Thus, we follow their suggestions and build a classifier to automatically calculate the set of probabilities Ψ for all the terms in W acting as facet terms in the posterior inference, i.e. $\psi_{d,n} \leftarrow g(\overrightarrow{X}, \rho)$ where $\overrightarrow{X}$ is a feature vector and ρ represents the learned classifier. X and ρ correspond to two observable variables in the plate notation of Figure 1(a). To build such a classifier, we resort to the TREC queries and the TREC judgments to collect typical topical terms and typical facet terms as training

[1] We only consider the terms appearing in at least 30 posts, because the average number of posts in the corpus for a query is about 21000. We believe that a term appearing in less than 30 posts can neither show up as a top topical term nor a facet term.

examples. The TREC judgments provide the personal and official feeds w.r.t. the TREC queries. We use query terms as typical topical terms, because query terms always express topical information. We propose a method called $STFT$ to select typical facet terms from the TREC judgments below.

1. For a TREC query q, partition the personal feeds from the official feeds w.r.t. q.
2. Pool the posts from the personal (official) feeds that contain q as the set of personal (official) posts P (O).
3. For each term in $P \cup O$, conduct the χ^2 test over P and O to calculate its χ^2 value; collect top-ranked terms (in descending order of their χ^2 values) as the candidates of typical facet terms for q.
4. Repeat steps 1 to 3 for each TREC query.
5. Pool the candidates for all TREC queries; for each candidate in the pool, record the number of queries of which it acts as a candidate.
6. Rank the candidates in the pool in descending order of that number from step 5 and select top terms as typical facet terms.

The high χ^2 value of a selected typical facet term shows that it has a clear facet inclination. The high number of queries of which a selected typical facet term appears as a candidate exhibits its strong generality. Thus we believe the $STFT$ method can select typical facet terms. For example, a typical personal term and a typical official facet term found by $STFT$ are "yeah" and "significantly" respectively.

The classification of topical and facet terms is challenging. Intuitively, typical topical terms tend to be nouns while typical facet terms are normally non-noun terms, such as pronouns and adverbs. However, our preliminary experimental results showed that poor performance was obtained if we just used the part-of-speech (POS) of a term to classify it, because there are exceptions. For example, the noun "product" can be an official facet term. To accurately separate topical terms from facet terms, we propose to use a term's POS and its contextual POSs as features, because these POSs of a term reflect its syntactic role. The syntactic role of a topical term and that of a facet term are likely to be different. Specifically, given a term $w_{d,n}$ in a post d, we first identify the sentence s containing $w_{d,n}$ and then obtain the POS of $w_{d,n}$, the POSs of 10 preceding terms of $w_{d,n}$ in s (if any) and those of 10 succeeding terms of $w_{d,n}$ in s (if any) as features.[2] This generates a vector of 21 features. Each topical (facet) training example consists of a typical topical (facet) term and a contextual sentence. We use as the contextual sentences the sentences containing at least one of the selected typical topical (facet) terms in the posts of the feeds from the TREC judgments.

After defining the features and collecting the training examples, we build a classifier that determines whether a term in a post acts as a topical term or a facet term. Specifically, given a term $w_{d,n}$ in a post d, if $w_{d,n}$ is classified to be a facet term with a class probability p (> 0.5), the probability of $w_{d,n}$ in d acting as a facet term is equal to p ($\psi_{d,n} = p$); if $w_{d,n}$ is classified to be a topical term with a class probability p' (> 0.5), the probability of $w_{d,n}$ in d acting as a facet term is equal to $1 - p'$ ($\psi_{d,n} = 1 - p'$). After we apply the classifier to all the terms in W, we obtain the set of probabilities $\Psi = \{\psi_{d,n} | d \in D; n = 1 \dots N_d\}$, each of which indicates the probability of a term in W acting as a facet term. Then we can use Ψ as a set of priors in the posterior inference (to be presented in Section 3.2). Note that the classifier does not specify whether a term in a post acts as a personal or an official term. A term's facet is inferred by the JTF model.

[2]In our preliminary experiments, the classification performance was improved with the increasing of the size of the contextual window (symmetrical at $w_{d,n}$) to collect the contextual POSs. The performance made negligible changes after the size of the window increased beyond 20.

3.2 Inference

The posterior inference predicts the topic (facet) distributions of posts Θ^T (Θ^F) and the term distributions of topics (facets) Φ^T (Φ^F). In order to infer these four distributions, we need to know the topic and facet assignments Z and the binary variables R for all the terms in W. Specifically, given a term $w_{d,n}$ in a post d, $z_{d,n}$ indicates the facet f (the topic t) assigned to $w_{d,n}$ if $r_d = 1$ ($r_d = 0$). We adopt the collapsed Gibbs sampling [3] to estimate the posterior distributions of Z and R. We skip the derivation details and estimate Z and R by the conditional probabilities below. [3]

$$P\left(z_{d,n} = f, r_{d,n} = 1 | Z^{-(d,n)}, R^{-(d,n)}, W^{-(d,n)}, w_{d,n} = v\right) \propto \psi_{d,n} \cdot \frac{\{m^F_{f,v}\}^{-(d,n)} + \beta_F}{\sum_{v' \in V}(\{m^F_{f,v'}\}^{-(d,n)} + \beta_F)} \cdot \frac{\{n^F_{d,f}\}^{-(d,n)} + \alpha_F}{\sum_{f' \in F}(\{n^F_{d,f'}\}^{-(d,n)} + \alpha_F)}$$

$$P\left(z_{d,n} = t, r_{d,n} = 0 | Z^{-(d,n)}, R^{-(d,n)}, W^{-(d,n)}, w_{d,n} = v\right) \propto (1 - \psi_{d,n}) \cdot \frac{\{m^T_{t,v}\}^{-(d,n)} + \beta_T}{\sum_{v' \in V}(\{m^T_{t,v'}\}^{-(d,n)} + \beta_T)} \cdot \frac{\{n^T_{d,t}\}^{-(d,n)} + \alpha_T}{\sum_{t' \in T}(\{n^T_{d,t'}\}^{-(d,n)} + \alpha_T)}$$

where $m^F_{f,v}$ and $m^T_{t,v}$ are the counts of the occurrences of the specific term $v (\in V)$ in W being assigned to the facet f and the topic t, respectively; $n^F_{d,f}$ and $n^T_{d,t}$ are the counts of the facet f and the topic t being assigned to the terms in d, respectively. The superscript $\{\cdot\}^{-(d,n)}$ means the exclusion of the term $w_{d,n}$. For example, $\{m^F_{f,v}\}^{-(d,n)}$ is the count of the occurrences of v being assigned to the facet f by excluding $w_{d,n}$ ($w_{d,n} = v$). The same applies to $Z^{-(d,n)}$, $R^{-(d,n)}$, $W^{-(d,n)}$, $\{m^T_{t,v}\}^{-(d,n)}$, $\{n^F_{d,f}\}^{-(d,n)}$ and $\{n^T_{d,t}\}^{-(d,n)}$. We provide some explanations for the two probabilities above. During the iterative sampling, the probability of a term $w_{d,n}$ in a post d being assigned to a facet f (a topic t) is proportional to the product of three items below:

1. The probability of $w_{d,n}$ in a post d acting as a facet (topical) term, $\psi_{d,n}$ ($1 - \psi_{d,n}$).
2. The smoothed ratio of the count of the occurrences of v in W excluding $w_{d,n}$ that are assigned to the facet f (the topic t) over the count of all the terms in W excluding $w_{d,n}$ that are assigned to the facet f (the topic t), $\frac{\{m^F_{f,v}\}^{-(d,n)} + \beta_F}{\sum_{v' \in V}(\{m^F_{f,v'}\}^{-(d,n)} + \beta_F)}$ $\left(\frac{\{m^T_{t,v}\}^{-(d,n)} + \beta_T}{\sum_{v' \in V}(\{m^T_{t,v'}\}^{-(d,n)} + \beta_T)}\right)$.
3. The smoothed ratio of the count of the terms in d excluding $w_{d,n}$ that are assigned to the facet f (the topic t) over the count of all the terms in d excluding $w_{d,n}$ that are assigned to facets (topics), $\frac{\{n^F_{d,f}\}^{-(d,n)} + \alpha_F}{\sum_{f' \in F}(\{n^F_{d,f'}\}^{-(d,n)} + \alpha_F)}$ $\left(\frac{\{n^T_{d,t}\}^{-(d,n)} + \alpha_T}{\sum_{t' \in T}(\{n^T_{d,t'}\}^{-(d,n)} + \alpha_T)}\right)$.

After the posterior distributions of Z and R are estimated, we estimate $\hat{m}^F_{f,v}$ and $\hat{n}^F_{d,f}$. Specifically, after we know all the topic or facet assignments to all the terms in W (indicated by Z and R), we estimate the count of the occurrences of

[3]For convenience of presentation, we omit some priors in the conditional probabilities in the paper. The omitted priors are $\beta_F, \beta_T, \alpha_F, \alpha_T$ and Ψ, i.e. $P(X|Y) = P(X|Y, \beta_F, \beta_T, \alpha_F, \alpha_T, \Psi)$.

$v(\in V)$ in W that are assigned to the facet f, $\hat{m}^F_{f,v}$, and the count of the terms in d that are assigned to the facet f, $\hat{n}^F_{d,f}$. Then we can estimate the facet distributions of posts Θ^F and the term distributions of facets Φ^F as follow: $\hat{\theta}^F_{d,f} = \frac{\hat{n}^F_{d,f}+\alpha_F}{\sum_{f'\in F}\left(\hat{n}^F_{d,f'}+\alpha_F\right)}$; $\hat{\phi}^F_{f,v} = \frac{\hat{m}^F_{f,v}+\beta_F}{\sum_{v'\in V}\left(\hat{m}^F_{f,v'}+\beta_F\right)}$. We can estimate $\hat{m}^T_{t,v}$ and $\hat{n}^T_{d,t}$, the topic distributions of posts Θ^T and the term distributions of topics Φ^T in a similar manner.

3.3 Personal and Official Classification

Given a corpus of posts, D, after the JTF model calculates the facet distribution of posts Θ^F and the term distributions of the two latent facets Φ^F, we can classify the posts in D into personal or official ones as follows.

1. Given the term distributions for the two latent facets, f_1 and f_2, we use the seed term "I" to determine which one of the two facets corresponds to the personal facet. Let Φ_{f_1} and Φ_{f_2} be the term distributions of f_1 and f_2 respectively. We compare the two probabilities of the term "I" for f_1 and f_2 ($\phi_{f_1,I}$ and $\phi_{f_2,I}$) to find the facet (say f_1) having a higher probability for "I". Then f_1 (f_2) corresponds to the personal (official) facet.
2. For each post $d \in D$, the probabilistic distribution of facets in d consists of two probabilities, θ_{d,f_1} and θ_{d,f_2} ($\theta_{d,f_1} + \theta_{d,f_2} = 1$). If $\theta_{d,f_1} > \theta_{d,f_2}$, d is classified to be a personal post with a personal facet score θ_{d,f_1}; if $\theta_{d,f_1} < \theta_{d,f_2}$, d is classified to be an official post with an official facet score θ_{d,f_2}.

4. EXTENDED JOINT TOPIC FACET MODEL

We now present an extended joint topic facet model called E-JTF. We observe that the posts from a feed are likely to show the same facet. The E-JTF model improves the JTF model by incorporating such an observation into its generative process. Let $\Lambda^F = \{\lambda^F_{e,f}\}_{|E|\times|F|}$ be the matrix where $\lambda^F_{e,f}$ is the probability of any post in the feed e exhibiting the facet f and the row λ^F_e is the probabilistic distribution of any post in e exhibiting all facets. The E-JTF model employs the feed affiliations of posts. Let the feed containing a post d be e_d. Now we present the generative process of the E-JTF model (see the plate annotation in Figure 1(b)).

1. For each facet $f \in F$, draw $\phi^F_f \sim Dirichlet(\beta_F)$;
2. For each topic $t \in T$, draw $\phi^T_t \sim Dirichlet(\beta_T)$;
3. For each feed $e \in E$, draw $\lambda^F_e \sim Dirichlet(\alpha_F)$;
4. For each post $d \in D$,
   1) Set $\theta^F_d = \lambda^F_{e_d}$, if $d \in e_d$;
   2) Draw $\theta^T_d \sim Dirichlet(\alpha_T)$;
   3) For each term in d, say $w_{d,n}$,
      i) Set $\psi_{d,n} \leftarrow g(\cdot)$;
      ii) Draw $r_{d,n} \sim Bernoulli(\psi_{d.n})$;
      iii) if $r_{d,n} = 0$ // *$w_{d,n}$ acts as a topical term.*
         a) Draw $z_{d,n} \sim Multinomial(\theta^T_d)$;
         b) Draw $w_{d,n} \sim Multinomial(\phi^T_{z_{d,n}})$;
      iv) else // *$w_{d,n}$ acts as a facet term.*
         a) Draw $z_{d,n} \sim Multinomial(\theta^F_d)$;
         b) Draw $w_{d,n} \sim Multinomial(\phi^F_{z_{d,n}})$;

The E-JTF model differs from the JTF model in that it enforces the same probabilistic distribution of exhibiting all facets to all the posts from a feed in its generative process (see step 4.1). We still draw a probabilistic distribution of exhibiting all topics for each post, because the posts from a feed are composed by a blogger at different moments and he/she may write about any topic. Intuitively, there may not be topical similarities among all the posts within a feed. The E-JTF model also involves the set of probabilities Ψ, each of which indicates the probability of a term in W acting as a facet term. We still utilize the classifier $g(\overrightarrow{X}, \rho)$ to learn Ψ in the posterior inference (see Section 3.1). During the posterior inference, we want to obtain the topic distributions of posts Θ^T, the facet distributions of feeds Λ^F and the term distributions of topics (facets) Φ^T (Φ^F). To obtain these four distributions, we need to know the topic or facet assignments Z and the binary variables R for all the terms in W. Again we leave out the derivation details and estimate the posterior distributions of Z and R by the collapsed Gibbs sampling as below.

$$P\left(z_{d,n} = f, r_{d,n} = 1 | Z^{-(d,n)}, R^{-(d,n)}, W^{-(d,n)}, w_{d,n} = v\right) \propto$$
$$\psi_{d,n} \cdot \frac{\{m^F_{f,v}\}^{-(d,n)}+\beta_F}{\sum_{v'\in V}(\{m^F_{f,v'}\}^{-(d,n)}+\beta_F)} \cdot \frac{\left(\sum_{d'\in e_d}\{n^F_{d',f}\}^{-(d,n)}\right)+\alpha_F}{\sum_{f'\in F}\left(\left(\sum_{d'\in e_d}\{n^F_{d',f'}\}^{-(d,n)}\right)+\alpha_F\right)}$$
$$P\left(z_{d,n} = t, r_{d,n} = 0 | Z^{-(d,n)}, R^{-(d,n)}, W^{-(d,n)}, w_{d,n} = v\right) \propto$$
$$(1-\psi_{d,n}) \cdot \frac{\{m^T_{t,v}\}^{-(d,n)}+\beta_T}{\sum_{v'\in V}(\{m^T_{t,v'}\}^{-(d,n)}+\beta_T)} \cdot \frac{\{n^T_{d,t}\}^{-(d,n)}+\alpha_T}{\sum_{t'\in T}(\{n^T_{d,t'}\}^{-(d,n)}+\alpha_T)}$$

We provide some explanations of the probabilities above. The probability of a term $w_{d,n}(= v \in V)$ in a post d being assigned to a topic t has the same interpretation as the JTF mode (see Section 3.2). The probability of a term $w_{d,n}$ in a post d of a feed e_d being assigned to a facet f is proportional to the product of three items:

1. The probability of $w_{d,n}$ acting as a facet term, $\psi_{d,n}$.
2. The smoothed ratio of the count of the occurrences of v in W excluding $w_{d,n}$ that are assigned to the facet f over the count of all the terms in W excluding $w_{d,n}$ that are assigned to the facet f, $\frac{\{m^F_{f,v}\}^{-(d,n)}+\beta_F}{\sum_{v'\in V}(\{m^F_{f,v'}\}^{-(d,n)}+\beta_F)}$.
3. The smoothed ratio of the count of the terms in the feed e_d excluding $w_{d,n}$ that are assigned to the facet f over the count of all the terms in e_d excluding $w_{d,n}$ that are assigned to facets, $\frac{\left(\sum_{d'\in e_d}\{n^F_{d',f}\}^{-(d,n)}\right)+\alpha_F}{\sum_{f'\in F}\left(\left(\sum_{d'\in e_d}\{n^F_{d',f'}\}^{-(d,n)}\right)+\alpha_F\right)}$.

The probability of a term $w_{d,n}$ in a post d being assigned to a facet f in the E-JTF model is calculated differently from that in the JTF model. Such probabilities in both models are the products of three items where their items 3 are different. Specifically, the item 3 for the JTF model is the estimated facet distribution within a post d by the terms in d assigned to facets in previous iterations. However, the item 3 for the E-JTF model is the estimated facet distribution within a feed e by the terms in e assigned to facets in previous iterations. Note that the item 3 for the E-JTF model is due to the generative process of the E-JTF model where all the posts from a feed are given the same facet distribution. This results in that the estimated facet distributions of all the posts from a feed are similar.

After finishing sampling Z and R, we can obtain the facet distributions of posts Θ^F and the term distributions for the two latent facets Φ^F by the same way as the one in the JTF model (see Section 3.2). Then we can classify the posts by the same way as the JTF model (see Section 3.3) by using Θ^F and Φ^F from the E-JTF model.

5. EXPERIMENTAL EVALUATION

Experimental Setups. We evaluate our models by using 8 TREC 2009 queries and 10 TREC 2010 queries that are

required to be searched over the TREC Blogs08 collection. TREC Blogs08 collection is the only collection available for personal and official blog distillation. Each query is associated with a personal facet and an official facet. Given a query q, a faceted blog distillation method is required to retrieve two rankings of feeds w.r.t. q, one ranking for each facet. The performance is evaluated by the TREC judgments for those queries. We evaluate the JTF and E-JTF models over three TREC baselines of feeds. Specifically, for each query q, we obtain a corpus of posts by pooling all the posts of all the unique feeds from the three TREC baselines w.r.t. q. We denote as *TREC query corpus* the corpus of posts for q. This produces 18 TREC query corpora. We apply the JTF model and the E-JTF model to a TREC query corpus, respectively. A model is effective and robust if its faceted performance constantly and significantly outperforms those of the three baselines. For both models, we set their priors $\alpha_T = \frac{50}{|T|}$, $\alpha_F = \frac{50}{|F|}$,$\beta_T = \beta_F = 0.1$ as suggested in [3]. We set the number of topics $|T| = 100$ and the number of facets $|F| = 2$ for each TREC query corpus and run the samplers for both models for 1000 iterations. We employ the mean average precision (MAP), the R-precision (R-pref), the normalized discounted cumulative gain over all positions (NDCG) and the precision at top 10 posts (P10) as the evaluation measures. MAP is most important [10]. The personal (official) MAP measure means the MAP measure in terms of the personal (official) performance. The same applies to R-prec, NDCG and P10 too.

Experimental Evaluation. We first qualitatively evaluate the JTF and E-JTF models. Specifically, we present the top (most representative) facet terms identified by both models. Due to space limit, we only present the facet terms identified by both models over the TREC query corpus w.r.t. an exemplified query, "*drug safety*". Table 1 shows the top facet terms identified by both models. We make bold some terms that are errors. Both models identify the representative personal facet terms that consist of two categorizes: 1) the first person pronouns, such as "*I*" and the interjection "*well*", and 2) some (simple) verbs, such as "*think*" and "*like*" that are frequently used in personal posts. Intuitively, these representative personal facet terms are rarely used in official posts. However, "*medical*" is an adjective related to "*drug safety*", but our models erroneously identify it as a personal facet term. Both models also identify the representative official terms. These terms consist of four kinds: 1) some adjectives, such as "*effective*"; 2) some adverbs, such as "*potentially*"; 3) some nouns, such as "*company*" and "*market*"; and 4) some verbs, such as "*develop*" and "*report*". Intuitively, these terms are more likely to be used in official posts than in personal posts. However, both models also identify some terms erroneously, some adjectives, such as "*pharmaceutical*", and some nouns, such as "*FDA*". These terms are related to the query, not general official facet terms.

We then evaluate the classifier that determines whether a term in a post acts as a topical term or a facet term. In practice, we use the terms from the TREC queries as the typical topical terms. Since we only have 18 queries, we can only collect 37 typical topical terms. After obtaining these typical topical terms, we collect the contextual sentences for them. Specifically, given a query q, we get all the posts that satisfy the following two conditions: 1) they are from the personal or official feeds w.r.t. q indicated by the TREC judgments; 2) they contain q. The posts satisfying these

	Top Facet Terms By the JTF Model
Personal	i, my, like, us, make, me, say, said, need, take, think, well, good, know, **medical**, our, see, want
Official	**clinical**, **medical**, current, **patient**, company, **million**, study, market, **pharmaceutical**, report, potentially, provide, develop, **fda**, product

	Top Facet Terms By the E-JTF Model
Personal	i, my, like, said, us, make, me, **medical**, say, think, well, know, take, good, our, need, want, see
Official	**clinical**, current, **medical**, company, **patient**, market, **pharmaceutical**, develop, potentially, provide, **million**, product, study, **additive**, effective

Table 1: Top Facet Terms Identified by JTF and E-JTF over TREC Query Corpus w.r.t. "drug safety".

	Precision	Recall	F1-Measure
Facet Class	0.76	0.72	0.74
Topical Class	0.73	0.77	0.75

Table 2: Facet and Topical Term Classification.

two conditions are likely to be relevant to q and the terms of q in such posts are likely to be topical terms. There are about 110K sentences from these selected posts that contain at least one term of q as the topical training examples. We employ the proposed $STFT$ method (see Section 3.1). The $STFT$ method selects 1021 typical facet terms. We randomly select about 110K sentences containing these typical facet terms as the facet training examples. We deliberately keep the training data balanced. We use the decision tree classifier in the Weka package[4] and conduct a 10-fold cross validation over the training data. Table 2 shows the average classification performance in Precision, Recall and F1-measure. Our classifier can separate topical terms from facet terms with a reasonable accuracy.

Note that we use the query terms as typical topical terms for training, so we build two classifiers and use them in the sequential experiments. Specifically, we test the TREC 2010 queries by using the classifier that is trained over the training examples for the TREC 2009 queries and vice versa.

Now we evaluate the faceted performance of both models. Specifically, we re-rank the feeds from each baseline by addressing their topical relevance to the queries and their extents of exhibiting a facet. This re-ranking process produces two rankings of feeds, one for the personal facet and the other for the official facet. To rank feeds by each model, we first calculate the facet scores of posts in feeds by that model and then aggregate the facet scores of posts to those of feeds. We adopt the method proposed in [5] to obtain the facet scores of feeds. An aggregated score of a feed f, $AS_k(f)$, combines the IR score of f, $IR(f)$ and its facet score (for a facet k) of f, $F_k(f)$: $AS_k(f) = \mu \cdot IR(f) + (1-\mu) \cdot F_k(f)$ where $IR(f)$ is provided by the TREC baselines and the parameter μ is empirically learned in the following manner. The specific value of μ that optimizes the faceted ranking performance for the TREC 2009 queries are used for the TREC 2010 queries and vice versa. All feeds are ranked in descending order of their aggregated scores.

We also compare both models with the LDA model [1] over the three baselines. Specifically, for each baseline, we use the LDA model to re-rank the feeds from the baseline in descending order of their topical relevance to the TREC queries. Note that the LDA model cannot measure the facets of feeds. It just calculates the IR scores of posts w.r.t. queries. Please refer to [14] for the details of such a calculation. Then we aggregate the IR scores of feeds by those of their posts and re-rank the feeds. We set up the LDA model's priors: $\alpha_T = \frac{50}{|T|}$, $\beta_T = 0.1$ as in [3] and $|T| = 100$.

[4]http://www.cs.waikato.ac.nz/ml/weka/

	Personal Facet Performance				Official Facet Performance			
	MAP	R-Prec	NDCG	P10	MAP	R-Prec	NDCG	P10
baseline1	0.2097	0.2234	0.4321	0.2222	0.2673	**0.2805**	0.4887	0.2222
LDA	0.1708†	0.2043	0.4149	0.1889	0.2307	0.2482	0.4472†	0.2222
JTF	0.2241‡	0.2244	0.4580‡	0.2222	0.2679	**0.2805**	0.4873‡	**0.2278**
E-JTF	**0.2731**‡	**0.2738**	**0.4956**‡	**0.2278**‡	**0.2742**‡	**0.2805**	**0.4919**‡	0.2222
baseline2	0.1527	0.1831	0.3455	0.1389	0.1957	0.1621	0.3945	0.1500
LDA	0.1615	0.1812	0.3672	0.1722	0.2502	**0.2818**†	**0.4612**†	**0.2222**†
JTF	0.1698	0.1884	0.3707	0.1722	0.2240	0.2544†	0.4315	0.1778‡
E-JTF	**0.2363**‡↑	**0.2618**	**0.4314**†‡↑	**0.2167**†‡↑	**0.2516**	0.2812†	0.4582†	0.2000†
baseline3	0.0895	0.1003	0.2508	0.1000	0.2016	0.1891	0.3658	0.1778
LDA	0.1295	0.1577	0.2988	0.1500†	0.2098	0.2235	0.3804	0.1778
JTF	0.1815	0.2097	0.3398	0.1389	0.2395	**0.2478**†	**0.4069**	**0.2056**†
E-JTF	**0.1881**†	**0.2345**†	**0.3426**	**0.1611**†	**0.2427**	0.2376	0.4057	**0.2056**†

Table 3: Personal and Official Performance of LDA, JTF and E-JTF over TREC Baselines. †, ‡ and ↑ indicate statistically significant changes over TREC baselines, LDA and JTF by one-sided paired t-test with $p < 0.05$.

	baseline1		baseline2		baseline3	
	Personal	Official	Personal	Official	Personal	Official
hitFeeds	0.2126	0.2700	0.1533	0.1957	0.0911	0.1985
LexMIRuns	0.2727	0.2662	0.1607	0.1882	0.0875	0.2016
QIOPFT	0.2440	0.2690	0.1966	0.2449	0.1683	0.2366
E-JTF	**0.2731**	**0.2742**	**0.2363**†‡	**0.2516**	**0.1881**†‡	**0.2427**†

Table 4: Personal and Official Performance of E-JTF vs. three State-of-the-Art Methods. † and ‡ indicate statistically significant improvements over hitFeeds and LexMIRuns by one-sided paired t-test with $p < 0.05$.

Table 3 shows the faceted performance of the LDA model, the JTF model and the E-JTF model over three TREC baselines (namely baselines1-3). Several observations are made. First, both the JTF and E-JTF models constantly outperform the three TREC baselines in the mean faceted performance, which shows them robust and effective. Second, the JTF model shows significantly better performance than the LDA model in the mean personal performance. It shows decent improvements over the LDA model in the mean official MAP, R-prec and NDCG and a slightly deterioration in the mean official P10 performance. Third, the E-JTF model displays significantly better performance than the LDA model in almost all the mean faceted performance except that a slight improvement over the LDA model in the mean official P10. These two observations show that both models are more effective than the LDA model. Four, the E-JTF model outperforms the JTF model in the mean faceted performance. The superiority of the E-JTF model to the JTF model validates the observation that all the posts from a feed are likely to have the same facet.

We also compare our best results achieved by the E-JTF model with the results achieved by the state-of-the-art methods. Specifically, we compare our E-JTF model with the "hitFeeds" runs [15], the "LexMIRuns" runs [7] and the "QIOPFT" runs [5]. All their results are reported by using the same testing benchmarks as ours: both TREC 2009 and TREC 2010 queries over the three TREC baselines. We compare our best results obtained by the E-JTF model with their results in terms of both faceted MAP performance. Table 4 shows the comparison of our results with the best-known results obtained by the three methods described above. We observe that the E-JTF model consistently and significantly outperforms all the best-known results in both faceted performance over all three baselines. These improvements demonstrate our E-JTF model is robust and effective.

6. CONCLUSION

We proposed two models that discover the topics and the personal and official facets of blog posts. Both models are supplemented by a classifier that separates topical terms from facet terms in posts during the posterior inference. Moreover, we observed and validated an important characteristic for personal and official blog distillation that all the posts from a feed are likely to exhibit the same facet. One of our two models considers such a characteristic in its generative process. We evaluated both models by the TREC 2009 and TREC 2010 queries over the TREC Blogs08 collection. Experimental results demonstrated the effectiveness of both proposed models. The results obtained by our second model outperform the best-known results.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[2] S. Gerani, M. Keikha, M. Carman, and F. Crestani. Personal blog retrieval using opinion features. ECIR'11, 2011.

[3] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1), 2004.

[4] L. Guo, F. Zhai, Y. Shao, and X. Wan. Pkutm at trec 2010 blog track. In *TREC*, 2010.

[5] L. Jia, C. Yu, and W. Meng. Facet-Driven Blog Feed Retrieval. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*, CICLing'13, 2013.

[6] L. Jia and C. T. Yu. Uic at trec 2010 faceted blog distillation. In *TREC*, 2010.

[7] M. Keikha, P. Mahdabi, S. Gerani, G. Inches, J. Parapar, M. J. Carman, and F. Crestani. University of lugano at trec 2010. In *TREC*, 2010.

[8] S. Li, Y. Li, J. Zhang, J. Guan, X. Sun, W. Xu, G. Chen, and J. Guo. Pris at trec 2010 blog track: Faceted blog distillaton. In *TREC*, 2010.

[9] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the trec 2009 blog track. In *TREC*, 2009.

[10] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the trec-2010 blog track. In *TREC*, 2010.

[11] Y. Mejova, V. Ha-Thuc, S. Foster, C. G. Harris, R. J. Arens, and P. Srinivasan. Trec blog and trec chem: A view from the corn fields. In *TREC*, 2009.

[12] A. Mukherjee and B. Liu. Aspect extraction through semi-supervised modeling. ACL'12, pages 339–348, 2012.

[13] R. L. T. Santos, R. M. C. McCreadie, C. Macdonald, and I. Ounis. University of glasgow at trec 2010: Experiments with terrier in blog and web tracks. TREC, 2010.

[14] M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. 2006.

[15] J. Yang, X. Dong, Y. Guan, C. Huang, and S. Wang. Hit_ltrc at trec 2010 blog track: Faceted blog distillation. TREC'10.

[16] W. X. Zhao, J. Jiang, H. Yan, and X. Li. Jointly modeling aspects and opinions with a maxent-lda hybrid. EMNLP'10.

[17] Z. Zhou, X. Zhang, and P. Vines. Rmit at trec 2010 blog track: Faceted blog distillation task. In *TREC*, 2010.