Find Me Opinion Sources in Blogosphere: A Unified Framework for Opinionated Blog Feed Retrieval

Xueke Xu[1,2], Songbo Tan[1], Yue Liu[1], Xueqi Cheng[1], Zheng Lin[1,2], Jiafeng Guo[1]
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
2. Graduate School of Chinese Academy of Sciences, Beijing, China
{xuxueke, tansongbo}@software.ict.ac.cn, {cxq, liuyue, guojiafeng, linzheng }@ict.ac.cn

ABSTRACT

This paper aims to find blog feeds having a principal inclination towards making opinionated comments on the given topic, so that we can subscribe to them to track influential and interesting opinions in the blogosphere. One major challenge is assigning topic-related opinion scores to blog feeds, which is embodied in two aspects. Firstly, we should identify whether the blog feed has a principal opinionated inclination. This inclination should be collectively revealed by all posts of the feed. We should fully consider evidences from all the posts of the feed to identify salient information among many posts of the feed. Secondly, we should capture topic-related opinions in the blog feed while ignoring irrelevant opinions.

In this paper, we propose a unified framework for opinionated blog feed retrieval, which combines topic relevance and opinion scores with a generative model. Furthermore, we propose a language modeling approach to estimating opinion scores that is seamlessly integrated into the framework, where two language models, Topic-specific Opinion Model (TOM) and Topic-biased Feed Model (TFM), work collectively to reflect whether the blog feed shows a principal on-topic opinionated inclination. To estimate TFM, we propose a topic-biased random walk to exploit both content and structural information to capture topic-biased salient information in the feed. As for TOM estimation, we propose to use a generative mixture model with prior guidance to effectively capture topic-specific opinion expressing language usage. The conducted experiments in the context of the TREC 2009-2010 Blog Track show the effectiveness of our proposed approaches.

Categories and Subject Descriptors

H .3.3 [Information Search and Retrieval]: Retrieval Model

General Terms

Algorithms, Performance, Experimentation.

Keywords

opinionated blog feed retrieval, topic-related opinionatedness, mixture model, topic-biased random walk

1. INTRODUCTION

Nowadays, millions of bloggers are expressing their opinions about various topics, making blogosphere a major information source of public opinions. There has been a considerable amount of research on opinion retrieval from blogosphere, most of which takes blog posts as retrieval units [18, 11, 31, 19, 5, 32, 6, 4]. Individual blog posts, however, could only provide users with limited opinion pieces. Largely different from blog post, a blog feed can provide users with continuously updated information. Actually, each blog feed is associated with a blog site (or blog in short), and could refer to a stream of posts issued by the author of the blog site with time [21]. Considering such a scenario where a user wishes to track influential opinions in blogosphere about a specific topic such as the economic policy of Obama government, she/he may subscribe to blog feeds which are dedicated to issuing opinions about this topic. Aiming at this need, in this paper we take blog feeds as the retrieval units, and study the task of *Opinionated Blog Feed Retrieval* [27]. Our task aims to find blog feeds showing a principal inclination towards making opinionated comments on the given query topic. By subscribing to these top retrieved feeds with their RSS readers, users may easily track public opinions of interest in time.

According to the aim of the task, a relevant blog feed should satisfy following two criteria: 1) *Topic Relevance*: The blog feed should have a principal, recurring interest in the given topic [13]. This criterion is necessary since topic relevance is a good indicator of whether the opinions are indeed about the given topic [24]. 2) *Topic-related Opinionatedness*: The blog feed should show a principal inclination towards expressing opinions about the topic. Actually, our task can be considered as a particular type of the faceted blog distillation introduced by TREC 2009 Blog Track [13]. We focus on the "opinionated vs. factual" facet and only consider the first value (i.e., "opinionated"). Most approaches in TREC 2009-2010 Blog Track follow a two-stage framework [13, 20]. Firstly, they estimate topic relevance to produce a topic relevance baseline ranking regardless of the opinion features. Secondly, they estimate opinion scores and re-rank blog feeds by combining topic relevance and opinion scores using a heuristic manner.

This task turned out to be very challenging in TREC 2009-2010 Blog Track since many participating approaches failed to improve the underlying baseline rankings [13, 20]. We argue that there are two major challenges. The first challenge is assigning opinion scores to blog feeds to reflect whether the feeds show principal inclination to expressing opinions towards the topic. This challenge is mainly embodied in following two aspects.

- Firstly, we should determine whether the feeds show prevalence to opinionatedness, and a burst of opinions in few posts is not adequate. In other words, the opinionated inclination should be collectively revealed by all the posts published by the corresponding blogger, not few posts. Thus, we should fully consider evidences from the feed to capture

WSDM'12, February 8-12, 2012, Seattle, Washington, USA.

salient content information which is shared among many posts of the feed, so that we can better identify whether the feed has a principal inclination to opinionatedness. To this end, we could exploit extra structural information in the blog feed beyond the content, such as relationships among the posts and words. Existing approaches usually ignore this valuable structural information and only take blog feed as a large document of the concatenation of all its constituent posts or a bag of independent posts.

- Secondly, we should take into account the query topic to capture opinions really related to topic at the granularity of blog feed, while ignoring irrelevant opinions. Many existing approaches in TREC consider opinions independently of the topic. Some other approaches consider topic-relatedness of opinions but not in an extensible and theoretical manner. They usually use heuristic techniques adapted from those for blog posts, such as "Near" information based [7], which may not be necessarily appropriate for the blog feeds.

In general, existing approaches to opinion scores estimation in TREC usually adopt lexicon-based [9, 10] or classification-based techniques [15, 7, 33] mostly adapted from those for individual blog posts. These approaches cannot well exploit special features of blog feeds, such as structural information. And most of these approaches have not been shown to be effective in the TREC results [13, 20], largely duo to the significant differences in the retrieval units and the aims between our task and blog post opinion retrieval.

The second challenge is finding a principled way to combine opinion scores and topic relevance to produce a final ranking. Existing approaches mostly estimate topic relevance and opinion scores in two separate stages and use a heuristic way to combine them, typically a linear summation [7, 9]. These two-stage approaches usually cannot well explore interaction between two factors of query topic and opinion to better balance between two criteria of topic relevance and topic-related opinionatedness for the final ranking. Indeed, heuristically considering opinion scores to re-rank the initial topic relevance ranking often largely harm topic relevance of the final ranking, to the point that it even deteriorates the overall performance compared with considering only topic relevance [13, 20].

This paper seeks to address these challenges in a unified framework. To this end, we specially introduce a hidden variable O_Q to denote the language usage of opinion expressions towards the given query topic Q, and rank blog feeds according to their generation probability given the query Q and O_Q, $P(F \mid Q, O_Q)$. Based on this generative model, we develop a unified probabilistic framework to estimate and combine topic relevance and opinion scores for blog feeds. In this unified framework, opinion scores are estimated using a language modeling approach. This approach determines whether the feed shows a clear on-topic opinionated inclination by how well the salient content information in the feed (with biased to the topic) fits the language usage of topic-related opinion expressions. The language usage of topic-related opinion expressions is captured by a language model, called Topic-specific Opinion Model (TOM). And the topic-biased salient content information is captured by another language model, Topic-biased Feed Model (TFM).

Language modeling approach provides an extensible, theoretical manner to fully and flexibly exploit various evidences from the blog feed, including both content information and structural information, to determine whether the blog feed show prevalence to opinionatedness towards the topic. We could take flexible ways to estimate the involved models. Specifically, in this paper, we propose a topic-biased random walk on Topic-specific Feed Graph to estimate TFM, which exploits topic-biased mutual reinforcement chain among posts and words to capture topic-biased salient content in the feed. As for TOM estimation, we propose to use a generative mixture model with prior guidance to effectively capture topic-specific opinion expression language usage.

We conduct empirical experiments in the context of the TREC 2009-2010 Blog Track. The results verify the unified framework and the proposed approaches to estimating the TFM and TOM. The results also largely outperform the best results in TREC 2009 and TREC 2010, respectively.

To sum up, the major contributions of this paper are:

- We propose a unified framework to estimate and combine topic relevance and opinion scores for opinionated blog feed retrieval.
- We propose a language modeling approach to estimating opinion scores, which is seamlessly integrated into the framework.
- We propose to use a generative probabilistic mixture model with prior guidance to estimate TOM.
- We propose a topic-biased random walk to exploits both content and structural information in the feed to estimate TFM.
- We conduct experiments in the context of the TREC 2009-2010 Blog Track to verify our models.

2. RELATED WORK

Blog post opinion retrieval. There has been a considerable amount of research on opinion retrieval from blogosphere, and most work takes blog posts as retrieval units [18, 11, 31, 5, 19, 32, 6, 22, 17, 23, 4, 25]. Blog post opinion retrieval aims at finding blog posts that have opinions about a given query topic [18]. At the first glance, our task is similar to blog post opinion retrieval in that they both consider topic and opinion for ranking respective retrieval objects. However, our task is largely different from that task in both retrieval units and the aims. We should fully exploit evidences from all posts of the feed, not a single post, to make a judge. These differences make our task especially challenging and the techniques adapted from those for individual blog posts are not necessarily effective as the results in TREC demonstrated [13, 20, 7, 9].

Opinionated blog feed retrieval. Most approaches in TREC 2009-2010 Blog Track follow a two-stage framework similar to that for blog post opinion retrieval [13, 20, 7, 9]. In the first stage, they estimate topic relevance and produce a topic relevance baseline ranking regardless of the opinionatedness criterion. In the next stage, they estimate opinion scores and re-rank blog feeds by combining topic relevance and opinion scores with a heuristic manner.

As for first stage, there is extensive research on topical blog feed retrieval (also referred to as blog distillation in TREC). Basically, the approaches may fall into two kinds according to their underlying blog representation models [2]: Small Document (SD) Model and Large Document (LD) Model. In SD Model, blog feeds are considered as collections of their constituent posts. The key issue is how to aggregate the topic relevance evidences of individual posts to infer the feed's topic relevance. To this end, various models have been explored such as blogger model [1], voting model [12] and resource selection model [2], etc. On the other hand, LD model treats a blog as a large document which is the concatenation of all its constituent posts. Generally, a language model (LM) is used to represent this large document, and ranking can be based on any LM based IR approaches [2].

As for second stage, most approaches in TREC adopt lexicon-based [9, 10] or classification-based [15, 7, 33] techniques that are similar to those devised for blog posts. Among those approaches, many consider opinions independently of the query topic, while some consider topic-relatedness of opinions using heuristic techniques adapted from those for blog post, such as "Near" information based [7]. Finally, topic relevance and opinion scores are combined using a heuristic manner, typically a linear summation [7, 9]

Here we present two typical approaches for the second stage in TREC, both failing to provide consistent and significant performance improvements over the underlying topic relevance baselines as our approaches. Keikha *et al.* [9] compute opinion score for each retrieved feed by averaging the opinionated weight for each word in the blog feed. Finally, the final ranking scores are computed as a linear combination of topic relevance scores estimated in the first stage and opinion scores. Jia *et al.* [7] use a topic-dependent SVM classifier to classify sentences into either opinionated or factual, and next use "NEAR" operator to determine whether the opinionated sentences are related to the topic. Opinion score is aggregated over all opinionated and topically relevant sentences in the blog feed. This opinion scores estimation approach is essentially adapted from that of [31] proposed for blog post opinion retrieval by treating each blog feed as a large document of the concatenation of all its constituent posts. Finally, topically retrieved blog feeds are re-ranked by linearly combining topic relevance and opinion scores.

To the best of our knowledge, Jiang *et al.*' work [8] is the only published work except for that in TREC. It uses a topic-opinion mixture model, constructed by linearly interpolating topic relevance model with opinion relevance model, to rank blog feeds according to KL divergence. Compared to our approach, it is essentially a linear combination of two factors of topic and opinion without solid theoretical justification. Besides, it simply treats each blog feed as a big document without fully consider structural information in the feed. In fact, that approach only managed to improve a very weak baseline.

3. THE UNIFIED FRAMEWORK

In this section, we will present the proposed unified framework that aims to effectively rank blog feeds by their likelihood of fulfilling both two criteria of topic relevance and topic-related opinionatedness discussed in Section 1. In particular, we will discuss in details the opinion scores estimation component in the framework.

According to the generative model in traditional information retrieval area, topic relevance can be estimated by probability of generating the blog feed F given the query Q, $P(F|Q)$. To consider the topic-related opinionatedness criterion in our task, we introduce a topic-specific variable O_Q to denote language usage of opinion expressions towards the query topic Q. Following the traditional generative model framework, we rank blog feeds according to their generation probability given the original query Q and O_Q, $P(F|Q,O_Q)$.

Formally, we have:

$$P(F|Q,O_Q) \propto P(F,Q,O_Q) = P(F)P(Q|F)P(O_Q|Q,F) \qquad (1)$$

There are two major components in Equation (1): $P(F)P(Q|F)$ deals with topical relevance, while $P(O_Q|Q,F)$ deals with opinion scores. This equation provides a justifiable framework for combining topical relevance and opinion scores, which is naturally induced from a generative model with a solid probabilistic theoretical foundation. The framework fully considers the highly dependence of opinions on the topic to better balance between the two criteria of topic relevance and topic-related opinionatedness for the final ranking.

The topic relevance of the feed F is considered as query generation probability given the feed combined with the feed prior, $P(F)P(Q|F)$. This component is not focus of this paper, since it can be estimated by existing approaches to topical blog feed search (also referred to as blog distillation in TREC) which has been extensively studied [2].

We here focus on the opinion scores estimation component, where the opinion score of the feed F can be considered as the probability of O_Q given the query Q and F, $P(O_Q|Q,F)$.

We marginalize $P(O_Q|Q,F)$ across all words in the vocabulary:

$$P(O_Q|Q,F) = \sum_{w\in V} P(w|Q,F)P(O_Q|w,F,Q) \qquad (2)$$

where V is the vocabulary, w is the word in V. By assuming conditional independence between O_Q and (Q,F) given the word w, Equation (2) reduces to:

$$P(O_Q|Q,F) = \sum_{w\in V} P(w|Q,F)P(O_Q|w) \qquad (3)$$

By assuming the probability of each word w (i.e. $P(w)$) to be uniform, and eliminating $P(O_Q)$ that doesn't affect the ranking, we get the following equation:

$$P(O_Q|w) = \frac{P(w|O_Q)P(O_Q)}{P(w)} \propto P(w|O_Q) \qquad (4)$$

Plugging Equation (4) into Equation (3), we come to:

$$P(O_Q|Q,F) \propto \sum_{w\in V} P(w|Q,F)P(w|O_Q) \qquad (5)$$

According to Equation (5), opinion score of feed F is estimated by accumulating the relatedness of all words from the feed with expressing topic-relevant opinions (i.e. $\{P(w|O_Q)\}_{w\in V}$), weighted by the topic-sensitive prominence of each word w in the feed (i.e. $P(w|Q,F)$).

Equation (5) provides a language modeling approach to estimating opinion scores. Specifically, we use a language model (LM, i.e. a probability distribution over the vocabulary) to estimate the probability $P(w|O_Q)$. We call this LM as Topic-specific Opinion Model (TOM). Besides, for each blog feed F, we also use a LM to estimate $P(w|Q,F)$, called as Topic-biased Feed Model (TFM). TOM should be estimated to reflect language usage of topic-related opinion expressions which helps identify topic relevant opinions. And TFM should be estimated to capture salient content information in the blog feed with bias towards the topic, and consequently help reflect whether the feed shows principal inclination to expressing opinions about the topic. Our approach, intuitively speaking, determines whether the feed shows a clear on-topic opinionated inclination by how well the salient content information in the feed (with biased to the topic) fits the language usage of topic-related opinion expressions.

Language modeling approaches have been attracting much attention in IR area due to its solid statistical foundation and extensibility by leveraging various estimation approaches [9]. Our approach, therefore, provides an extensible, theoretical manner to fully and flexibly exploit various evidences from the blog feed, including both content information and structural information, to determine whether the blog feed show prevalence to opinionatedness towards the topic.

Note that, Zhang et al. [32] and Gerani *et al.* [4] have, respectively, proposed unified generation models for *blog post opinion retrieval*. The essential difference between their models and ours lies in how to deal with the challenge of opinion scores estimation. Their models use a general opinion word lexicon, and exploit the proximity based information (e.g. positional closeness of query terms to the general opinion words in the post documents) to capture the topic relevant opinions in the posts. However, it's non-trivial to adapt the proximity based approach for individual documents to blog feeds duo to their significant differences in granularity [7]. On the other hand, due to the specially introduced topic-specific opinion variable O_Q, our approach could fully exploit language usage information of opinion expressions about the topic to effectively capture relevant opinions in feeds [1]. More importantly, we could fully and flexibly exploit the special features of feeds for estimating their opinion scores under a language modeling framework.

Now, the key issue is to find a best way to estimate TOM and TFM, so that they can work together to reflect whether the blog feed shows prevalence to opinionatedness towards the topic. In following two sections, we will, respectively, discuss in details the requirements for better estimating TOM and TFM, and describe the estimation details.

4. TOM ESTIMATION

4.1 Requirements for Better Estimating TOM

TOM (i.e. $\{P(w \mid O_Q)\}_{w \in V}$) is required to reflect the language usage of topic-relevant opinion expressions. In other word, TOM should capture opinion words frequently used to express opinions towards the given topic. Indeed, people tend to use different opinion words to express opinions for different topics. And topic-specific opinion words are naturally more indicative of an opinion that is really towards the topic than other opinion words [4]. For instance, the word "rhythmic" may be used more for expressing opinion about music related topic than other topics, and it is in turn more indicative of an opinion about music than general opinion words like "great". Thus, we should assign a high probability value to a topic-specific opinion word, which is more likely to indicate a relevant opinion, and assign a relatively low value to a topic-unrelated general opinion word, which usually indicates an irrelevant opinion, and factual words, which doesn't indicate an opinion. In this way, TOM could help identify topic-related opinions, and ignore irrelevant opinions.

4.2 Mixture Model

In the context of opinion retrieval task, existing approaches usually learn a TOM by separately weighting words based on pseudo opinion relevance feedback [6, 17, 8], where the learned TOM could be easily "contaminated" by highly frequent, non-discriminative words or factual topic-related words. We here instead use a generative probabilistic mixture model, which could model topic-specific opinion expressions in a more effective and principled way. In particular, a background model that reflects general information in the background collection is used to prevent the learned model from being contaminated by general and usually topic-unrelated words and make the learned model more discriminative. Besides, we introduce a prior for TOM to discriminate opinionated content from factual content and prevent the learned model from being contaminated by factual topic-related words. Various variants of mixture models have been widely applied to different text analysis tasks [30, 16], we here extend the application of mixture models to modeling topic specific opinions in the context of opinion retrieval task.

Our model could be roughly considered as a simplification version of Topic Sentiment Mixture model [16], where only one topic and one sentiment are involved. Specifically, in our mixture model, the words in a topically relevant post are assumed to be generated by sampling from a mixture model involving TOM, Topic Relevance Model (TRM) and the background model. Formally, let θ_O be the TOM, θ_F be the TRM and θ_B be the background model. The generation likelihood of word w in post d is given as:

$$P_d(w) = \lambda_B P(w \mid \theta_B) + (1-\lambda_B)(\pi_{d,O} P(w \mid \theta_O) + \pi_{d,F} P(w \mid \theta_F))$$

where λ_B is a fixed weight controlling the influence of the background model, $\pi_{d,O}$ ($\pi_{d,F}$) is the mixing weight of θ_O (θ_F) for post d, and $\pi_{d,F} + \pi_{d,O} = 1$.

Let $\boldsymbol{C} = \{d_1, d_2, ..., d_m\}$ be a set of topically relevant posts, then the generation log-likelihood of $\boldsymbol{C}$ is given as:

$$\log P(\boldsymbol{C} \mid \Lambda) = \sum_{d \in \boldsymbol{C}} \sum_{w \in V} [c(w,d) \times \log(\lambda_B P(w \mid \theta_B) + (1-\lambda_B)(\pi_{d,O} P(w \mid \theta_O) + \pi_{d,F} P(w \mid \theta_F)))]$$

where $\Lambda = \{\theta_O, \theta_P, \pi_{d,O}, \pi_{d,F}\}$ is the parameter set to estimate, V is the vocabulary, $c(w,d)$ is the frequency of word w in document d.

We set λ_B to 0.95 as Zhai *et al.* suggested [30] to alleviate the influence of general-purpose words and make the learned TOM more discriminative. Note that, θ_B is estimated beforehand using Maximum Likelihood Estimation (MLE) based on the whole Blogs08 collection [14] and will be fixed during the learning process. To obtain $\boldsymbol{C}$, we use the original query terms to retrieve the top 500 topically relevant posts using the BM25 model from the Blogs08 collection.

We use Expectation-Maximization (EM) algorithm to compute a Maximum Likelihood Estimation of Λ as following updating formulas:

E-steps:

$$P(z(d,w,O)) = \frac{\pi_{d,O}^{(n)} P^{(n)}(w \mid \theta_O)}{\sum_{\upsilon \in \{O,F\}} \pi_{d,\upsilon}^{(n)} P^{(n)}(w \mid \theta_\upsilon)}, \; P(z(d,w,F)) = 1 - P(z(d,w,O))$$

$$P(z(d,w,B)) = \frac{\lambda_B P(w \mid \theta_B)}{\lambda_B P(w \mid \theta_B) + (1-\lambda_B) \sum_{\upsilon \in \{O,F\}} \pi_{d,\upsilon}^{(n)} P^{(n)}(w \mid \theta_\upsilon)}$$

M-steps:

$$\pi_{d,O}^{(n+1)} = \frac{\sum_{w \in V} c(w,d) P(z(d,w,O))}{\sum_{\upsilon \in \{O,F\}} \sum_{w \in V} c(w,d) P(z(d,w,\upsilon))}, \; \pi_{d,F}^{(n+1)} = 1 - \pi_{d,O}^{(n+1)}$$

$$P^{(n+1)}(w \mid \theta_O) = \frac{\sum_{d \in \pounds} c(w,d)(1 - P(z(d,w,B))) P(z(d,w,O))}{\sum_{w \in V} \sum_{d \in \pounds} c(w,d)(1 - P(z(d,w,B))) P(z(d,w,O))}$$

$$P^{(n+1)}(w \mid \theta_F) = \frac{\sum_{d \in \pounds} c(w,d)(1 - P(z(d,w,B))) P(z(d,w,F))}{\sum_{w \in V} \sum_{d \in \pounds} c(w,d)(1 - P(z(d,w,B))) P(z(d,w,F))}$$

[1] Note that, we could also additionally use proximity based evidences in estimating TFM to help further capture topic-related opinions. And it will be very interesting in future to develop an appropriate way to exploiting such evidences for blog feeds. In this paper, we use "topic bias" in random walk in estimating TFM to help further capture topic-related opinions

where $z(d,w,O)$ ($z(d,w,F)$) is a hidden variable which denotes that word w in document d is generated from θ_O (θ_F).

Without any prior knowledge as guidance, the learned θ_O and θ_F can not differentiate with each other because opinionated content and factual content generally co-occur with each other even in a highly opinionated post. Thus the estimated θ_O would be biased towards factual topic-related words, and thus can't effectively reflect the characteristics of topic-related opinion expressions. To address this problem, we introduce a general opinion model $\bar{\theta}_O$ as prior knowledge for TOM to discriminate opinions from factual content. $\bar{\theta}_O$ is derived based on a general opinion lexicon (denoted as GO) with each opinion word in the lexicon uniformly distributed as:

$$P(w \mid \bar{\theta}_O) = \begin{cases} \frac{1}{|GO|}, if\ w \in GO \\ 0 \end{cases}$$

The general opinion lexicon is built based on two publicly available opinion knowledge bases. We first select from SentiWordNet [3] a list of words with a positive or negative score above a given threshold (i.e. 0.6). Then we extract from MPQA subjectivity lexicon[2] another list of words with corresponding type being "strongsubj". Finally we construct the opinion lexicon as union of these two word lists. To incorporate the prior knowledge, we define a conjugate Dirichlet prior for θ_O : $Dir(\{1+\mu P(w \mid \bar{\theta}_O)\}_{w\in V})$, and use uniform prior for other parameters, then the prior of all parameters is as:

$$P(\Lambda) \propto P(\theta_O) = \prod_{w\in V} P(w|\theta_O)^{\mu P(w|\bar{\theta}_O)}$$

The parameter μ indicates the confidence of the prior. With this prior, we can use Maximum A Posteriori estimation: $\tilde{\Lambda} = \arg\max_{\Lambda}(P(\Lambda)P(\boldsymbol{C} \mid \Lambda))$, and the corresponding updating formula for θ_O in M-steps of the above EM algorithm is modified as follows:

$$P^{(n+1)}(w \mid \theta_O) = \frac{\mu P(w \mid \bar{\theta}_O) + \sum_{d\in \pounds} c(w,d)(1-P(z(d,w,B)))P(z(d,w,O))}{\mu + \sum_{w\in V}\sum_{d\in \pounds} c(w,d)(1-P(z(d,w,B)))P(z(d,w,O))} \tag{6}$$

Intuitively, the impact of incorporating this prior is equivalent to adding $\mu p(w \mid \bar{\theta}_O)$ pseudo counts for word w in estimating θ_O . In this way, the opinion words frequently used within the will stand out in the learned TOM. For instance, for the TREC topic "jazz music", such words as "rhythmic", "melodic", "dreamy", and "superb" are among top words of the trained TOM.

5. TFM ESTIMATION

5.1 Requirements for Better Estimating TFM

TFM (i.e. $\{P(w \mid Q,F)\}_{w\in V}$) is estimated aiming to capture salient content information of the blog feed with bias towards the topic, and consequently, help better determine whether the blog feed is clearly inclined towards expressing opinions about the topic. We argue that, to this end, the specific requirements are embodied in following two aspects.

- *Salience.* We should capture salient content information shared among many posts of the feed, ignoring trivial content in only few posts, aiming to reflect whether the feed has a principal opinionated inclination. To this end, we should fully exploit all evidence from the feed, including content and structural information. Straightforwardly taking each feed as a large document of concatenation of its constituent posts is not appropriate. Because it would be easily biased to few long posts or influenced by trivial or noisy content in the feed, and cannot effectively reflect the principal inclination of the feed.
- *Topic Bias.* We should emphasize more topically relevant content in the feed, aiming to further capture relevant opinions and discard those irrelevant. Our assumption is that an opinion co-occurring with topically relevant content is more likely towards the target topic, and thus should be highlighted. On the other hand, an opinion within topically irrelevant content is less likely towards the target topic, and thus should be assigned with low weight.

5.2 Our Solution

Our proposed solution is to some degree inspired by [26], which exploits relationships among documents, sentences and words to identify topic-biased salient information within documents for topic-focused text summarization. Specifically, we use a Topic-specific Feed Graph to represent each feed under the specific query topic. We then propose a topic-biased random walk on the graph, which exploits topic-biased mutual reinforcement chain among posts and words. In this way, we consider the above two aspects of requirements simultaneously to balance between them for a better estimation of TFM.

5.2.1 Topic-specific Feed Graph

The graph includes two types of nodes representing posts and words respectively, and multiple types of edges corresponding to relationships among them. Specifically, give a feed F and the query Q, a weighted undirected graph is defined as: $\boldsymbol{G}^Q = (P, W, E^{PP}, E^{PW}, E^{WW}, \boldsymbol{M}^{Q,PP}, \boldsymbol{M}^{PW}, \boldsymbol{M}^{WW})$, where P is the node set of all posts of feed F, and W is the node set of all words in the feed. E^{PP} is the edge set between posts and posts, E^{PW} between posts and words, and E^{WW} between words and words. All edges are associated with weights to measure the relationships between the corresponding objects, and the weighting matrices for E^{PP} , E^{PW} and E^{WW} are $\boldsymbol{M}^{Q,PP}$, $\boldsymbol{M}^{PW}$ and $\boldsymbol{M}^{WW}$, respectively.

We use the matrix $\boldsymbol{M}^{Q,PP}$ to reflect weights on $\boldsymbol{E}^{PP}$ under query Q, where $M_{i,j}^{Q,PP}$ measures topic-sensitive similarity relationship between post p_i and p_j . Specifically, $M_{i,j}^{Q,PP}\ (i \neq j)$ is computed as:

$$M_{i,j}^{Q,PP} = \sum_{w\in V} \text{TF-IDF}(p_i,w) \cdot \text{Weight}(w,Q) \cdot \text{TF-IDF}(p_j,w) \tag{7}$$

here $\text{TF-IDF}(p,w)$ is L_2-normalized *TF-IDF* weight of word w in post p ,and $\text{Weight}(w,Q)$ indicates the relatedness of word w to the query topic. We use Bol model [5] to compute $\text{Weight}(w,Q)$, measuring how informative the word w is in a collection of pseudo relevance feedback posts[3] against the background collection (i.e. Blogs08 collection [14]). This equation highlights the contribution of topic-related words, and thus makes the similarity calculation topic sensitive. Note that, $M_{i,j}^{Q,PP}(i=j)$ is set to 0 to avoid self-reinforcement. Then $\boldsymbol{M}^{Q,PP}$ is normalized to $\tilde{\boldsymbol{M}}^{Q,PP}$ by making sum of each row equal to 1.

We use the matrix $\boldsymbol{M}^{WW}$ to reflect weights on E^{WW} , where $M_{i,j}^{WW}$ measures relationship between word w_i and word w_j in the feed. $M_{i,j}^{WW}\ (i \neq j)$ is computed as Pointwise Mutual Information between word w_i and w_j based on co-occurrence

[2] http://www.cs.pitt.edu/mpqa/

[3] In our experiments, we use original query terms to retrieve top 30 posts from Blogs08 collection using BM25 model.

statistics at sentence level in the feed as follow:

$$M_{i,j}^{WW} = \log\left(\frac{\mathrm{P}(w_i, w_j)}{\mathrm{P}(w_i)\times \mathrm{P}(w_j)}\right) = \log\left(\frac{\mathrm{count}(w_i, w_j)\times \mathrm{S}}{\mathrm{count}(w_i)\times \mathrm{count}(w_j)}\right)$$

where S is total number of sentences in the feed, $\mathrm{count}(w)$ is the count of sentences containing word w, and $\mathrm{count}(w_i, w_j)$ is the count of sentences containing both w_i and w_j .Note that, a very small number (i.e., $1/|W|$) is added to each count for smoothing. $M_{i,j}^{WW}\,(i=j)$ is set to 0 to avoid self-reinforcement. Then $\boldsymbol{M}^{WW}$ is normalized to $\tilde{\boldsymbol{M}}^{WW}$ by making the sum of each row equal to 1.

At last, we use the matrix $\boldsymbol{M}^{PW}$ to reflect weights on $\boldsymbol{E}^{PW}$, and $M_{i,j}^{PW}$ measures relationship between post p_i and word w_j . $M_{i,j}^{PW}$ is computed as $\text{TF-IDF}(p_i, w_j)$. We use another matrix $\boldsymbol{M}^{WP}$ to denote the transpose of $\boldsymbol{M}^{PW}$ .Then $\boldsymbol{M}^{PW}$ and $\boldsymbol{M}^{WP}$ are, respectively, normalized to $\tilde{\boldsymbol{M}}^{PW}$ and $\tilde{\boldsymbol{M}}^{WP}$ by making the sum of each row equal to 1.

5.2.2 Topic-biased Random Walk

We here propose a topic-biased random walk on the graph, which exploit topic-biased mutual reinforcement chain among posts and words to capture topic-biased salient content in the feed. The basic idea of mutual reinforcement principle is embodied in following assumptions.

1. A post is salient, if (1) it is similar to many other salient posts; (2) it contains many salient words.

2. A word is salient if (1) it is strongly associated with many other salient words; (2) it appears in many salient posts.

Besides the above assumptions, we will further consider "topic bias" to highlight more topically relevant content in the feed.

Specifically, let $\boldsymbol{R}_P = [R_P(p_i)]_{|P|\times 1}$ and $\boldsymbol{R}_W = [R_W(w_j)]_{|W|\times 1}$ denote salience score vectors for P and W, respectively. Then topic-biased mutual reinforcement principle can be encoded in following iterative equations.

According to assumption 1, the salience score computation for P is formulated in an iterative form as:

$$\begin{aligned} R_P(p_i) = {} & \lambda_{11}[\alpha \textstyle\sum_{j=1}^{|P|} \tilde{M}_{j,i}^{Q,PP} R_P(p_j) + (1\text{-}\alpha) A_P^Q(p_i) R_P(p_j)] \\ & + \lambda_{21}[\alpha \textstyle\sum_{j=1}^{|W|} \tilde{M}_{j,i}^{WP} R_W(w_j) + (1\text{-}\alpha) A_P^Q(p_i) R_W(w_j)] \end{aligned} \quad (8)$$

Likewise, according to assumption 2, the salience score computation for W is formulated in an iterative form as:

$$\begin{aligned} R_W(w_i) = {} & \lambda_{12}[\alpha \textstyle\sum_{j=1}^{|P|} \tilde{M}_{j,i}^{PW} R_P(p_j) + (1\text{-}\alpha) A_W(w_i) R_P(p_j)] \\ & + \lambda_{22}[\alpha \textstyle\sum_{j=1}^{|W|} \tilde{M}_{j,i}^{WW} R_W(w_j) + (1\text{-}\alpha) A_W(w_i) R_W(w_j)] \end{aligned} \quad (9)$$

where $\boldsymbol{A}_P^Q = [A_P^Q(p_i)]_{1\times|P|}$ and $\boldsymbol{A}_W = [A_W(w_i)]_{1\times|W|}$ is the preference probability vector for P and W, respectively. We use a uniform preference probability vector for W, but use a topic-biased preference probability vector for P, i.e. $A_P^Q(p_i) \propto \mathrm{BM25}(p_i, Q)$ [4], to favor topically relevant posts.

The parameters $\lambda_{lm}(l=1,2; m=1,2)$ control the relative importance of different types of relationships, and we have: $\lambda_{11}+\lambda_{12}=\lambda_{21}+\lambda_{22}=1$. In our experiments, we simply set all λ_{lm} to 0.5 so as to fully exploit all types of relationship information. The parameter α is empirically set to 0.85 as the PageRank.

The initial $R_P(p_i)$ is set to $1/|P|$ for each post, and the initial $R_W(w_i)$ is set to $1/|W|$ for each word, so that $\|\boldsymbol{R}_P\|_1 = \|\boldsymbol{R}_W\|_1 = 1$.

[4] We add a small value to each BM25 score so that the preference value for each post is larger than 0.

Then, the salience scores computation could be conducted by iteratively running the Equations (8-9). Note that, it could be easily checked that $\|\boldsymbol{R}_P\|_1$ and $\|\boldsymbol{R}_W\|_1$ will keep being 1 during iteration process under our parameter setting (i.e. λ_{lm} =0.5).

This iterative process could be considered as a topic-biased random walk on the feed graph, where the states are nodes of the graph and the transition matrix is given as:

$$\tilde{\tilde{\boldsymbol{M}}}^{Q} = \begin{bmatrix} \lambda_{11}\tilde{\tilde{\boldsymbol{M}}}^{Q,PP} & \lambda_{12}\tilde{\tilde{\boldsymbol{M}}}^{PW} \\ \lambda_{21}\tilde{\tilde{\boldsymbol{M}}}^{WP} & \lambda_{22}\tilde{\tilde{\boldsymbol{M}}}^{WW} \end{bmatrix}.$$

The block matrix $\tilde{\tilde{\boldsymbol{M}}}^{Q,PP}$ in $\tilde{\tilde{\boldsymbol{M}}}^{Q}$ corresponds to the local transition probability from posts to posts, and we have $\tilde{\tilde{\boldsymbol{M}}}^{Q,PP} = \alpha\tilde{\boldsymbol{M}}^{Q,PP} + (1-\alpha)[1]_{|P|\times 1}\cdot \boldsymbol{A}_P^Q$. Similarly, $\tilde{\tilde{\boldsymbol{M}}}^{PW}$ corresponds to the local transition probability from posts to words, and we have $\tilde{\tilde{\boldsymbol{M}}}^{PW} = \alpha\tilde{\boldsymbol{M}}^{PW} + (1-\alpha)[1]_{|P|\times 1}\cdot \boldsymbol{A}_W$. The other block matrices in $\tilde{\tilde{\boldsymbol{M}}}^{Q}$ are constructed likewise. Then, the iterative Equation (8-9) could be rewritten in a matrix form as:

$$\begin{bmatrix} \boldsymbol{R}_P \\ \boldsymbol{R}_W \end{bmatrix} = (\tilde{\tilde{\boldsymbol{M}}}^{Q})^T \cdot \begin{bmatrix} \boldsymbol{R}_P \\ \boldsymbol{R}_W \end{bmatrix}$$

It can be easily checked that the transition matrix is irreducible and aperiodic, thus a stationary score vector $\begin{bmatrix} \hat{\boldsymbol{R}}_P \\ \hat{\boldsymbol{R}}_W \end{bmatrix}$ can be obtained after adequate iterations [5]. Finally, we compute TFM as: $P(w_i \mid Q, F) \propto \hat{R}_W(w_i)$.

Note that, "topic biased" in this random walk is embodied in two aspects: 1) topic-sensitive similarity measure between posts as Equation (7), which makes the salience score transition sensitive to the topic to prevent topic-drift, and 2) topic-biased preference probability vector for P , which rewards those more topically relevant posts.

6. EXPERIMENTAL RESULTS

6.1 Experimental Setting

Test Collection. We conduct experiments in the context of faceted blog distillation task in TREC 2009-2010 Blog Track. In that task, each topic is associated with an additional "facet" field besides the traditional TREC topic fields. Each facet has two values, and each value corresponds to a separate ranking of blogs. An example topic is shown in Figure 1. Note that, we here focus on the "opinionated vs. factual" facet and only consider the first value (i.e. "opinionated").

There are totally 13 TREC 2009 topics and 7 TREC 2010 topics associated with "opinionated vs. factual" facet and officially used for evaluation. We use all these topics along with the corresponding official relevance judgments for test. The relevance judgments are in five scales [13], and we consider as being relevant the feeds that are topically relevant and clearly inclined towards first facet value (i.e. "opinionated") in the context of our task. Besides, we only use the "query" field of the topics as query terms discarding other fields such as "desc".

We use the TREC Blogs08 collection adopted in the TREC 2009-2010 Blog Track [14], which is a large scale of sample of blogosphere between 14/01/2008 and 10/02/2009. As for data preprocessing, we adopt a link tables removing algorithm [24] to detect valuable content blocks from post pages and discard noisy

[5] In our experiment, the iteration count is empirically set to 10.

blocks. We remove stop words from the extracted post content based on a stop word list[6], but not perform word stemming.

Topic Relevance Baselines. In order to facilitate fair comparisons among different opinion-based re-ranking techniques, TREC 2010 organizers selected three TREC standard baselines[7] from participating runs for baseline blog distillation task that considers only topic relevance as criterion [20]. Among these baselines, stabaseline1 is one of best performing runs for baseline blog distillation task in TREC 2010; stdbaseline3 could be treated as a weak run; while stdbaseline2 could represent a median run [20]. With TREC standard baselines in hand, we could evaluate the effectiveness and robustness of our proposed approaches by how much performance improvement over these baselines could be achieved. We adopt these standard baselines to implement topic relevance component in the unified framework (see Section 3). Specifically, we use the topic relevance scores provided by the corresponding baseline as topic relevance probability (i.e. $P(F)p(Q|F)$ in Equation (1)). Note that, more appropriate approaches to transforming topic relevance scores to topic relevance probability may further improve the performance [4], and this is a part of our future work.

<top>
<num> Number: 1162 </num>
<query> uzbekistan </query>
<desc> Description:
I am interested in news from Uzbekistan.
</desc>
<facet> opinionated </facet>
<narr> Narrative:
I am interested in information about what is happening in Uzbekistan (current events, not history). A blog that lists Uzbekistan in a list of countries is judged not relevant.
</narr>
</top>

Figure 1: Blog Track 2010, faceted blog distillation task, topic "1162".

Evaluation Metrics. The evaluation metrics we use are standard IR measures, such as mean average precision (MAP), R-Precision (R-prec), and precision at the top 10 results (p@10).

Approaches to Estimating TOM. Besides the approach presented in Section 4, referred to as **Mix,** we here introduce alternative approaches for comparisons. We will verify the reasonability of the requirements discussed in Section 4.1 for TOM estimation by comparing Mix with these additional approaches.

- **GEN.** This approach takes the general opinion model $\overline{\theta}_O$ (also used as prior in the mixture model, see Section 4) as TOM. This estimation is general across all topics, but cannot capture topic-specific characteristics of opinion expressions.
- **PRF.** The outline of this approach can be summarized in the following steps. Firstly, we use the original query terms to retrieve the top 5000 topically relevant posts from the Blogs08 collection with the BM25 model. Secondly we use all together the words in the opinion lexicon (also used in Section 4) as a query to re-retrieve top 30 posts from the 5000 posts as opinion relevance feedback posts. At last, we use divergence minimization algorithm [28] to estimate TOM, which weights each word by how discriminative the word is in opinion relevance feedback posts against the background collection (i.e. Blogs08 collection). This approach is quite similar to that used in [8], which also uses divergence minimization algorithm to estimate opinion relevance model based on opinion relevance feedback documents. The major limitation of this approach is that it can't effectively separate opinion words from factual topic-related word since they often co-occur with each other even in highly opinionated relevant posts. Thus the learned model may be biased towards factual topic-related words.

Approaches to Estimating TFM. Besides the approach presented in Section 5, which is referred to as **TRW,** we here further introduce alternative approaches. We will verify reasonability of the two aspects of requirements discussed in Section 5.1 for TFM estimation by comparing TRW with these additional approaches.

- **MLE**. It takes each feed as a large document of concatenation of all its constituent posts and uses Maximum Likelihood Estimation (MLE) with Jelinek-Mercer (JM) smoothing to estimate TFM. JM smoothing is more effective than other smoothing approaches for long and verbose queries according to Zhai & Lafferty's empirical study [29]. The smoothing parameter lambda is set empirically to 0.95. The limitation of this approach is that it would be easily influenced by trivial or noisy content in the feed, and cannot effectively reflect the salient information in the feed (i.e. not considering the first aspect of the requirements very well). Besides, it also ignores the second aspect of the requirements.
- **RW**. This approach is similar to TRW, but considering no "topic bias" by taking a uniform prior for *P,* and a non-topic-sensitive cosine similarity measure to compute $M_{i,j}^{Q,PP}$ (see Section 5). This approach can well capture the salient content information in the feed (i.e. considering the first aspect of the requirements), but ignore the second aspect.

Approaches to Estimating Opinion Score. An opinion score estimation approach could be a flexible combination of any TOM estimation approach and TFM estimation approach. Our proposed approach, referred to as Mix-TRW, uses Mix and TRW to estimate TOM and TFM, respectively. As comparisons, we introduce additional variants for our approach, which use different techniques to estimate the TOM and TFM. For instance, GEN-MLE refers to the approach using GEN and MLE to estimating TOM and TFM, respectively.

6.2 Results and Analysis

These are three groups of results in Table1, Table3 and Table4, respectively, each group corresponding to one of TREC standard baselines. For each group, we use the corresponding baseline to implement the topic relevance component in our unified framework (i.e. $P(F)P(Q|F)$ in Equation (1)). We will give comparisons among different TOM estimation approaches, as well as comparisons among different TFM estimation approaches. Through comparisons, we will verify the proposed requirements for an appropriate estimation of TOM and TFM, respectively (See Section 4.1, 5.1). And we will also show the effectiveness of approaches to estimating TOM and TFM presented in Section 4 and 5, respectively. Note that, the default value of parameter μ in Mix for estimating TOM (See Equation (6) in Section 4) is experimentally set to 100, 000.

Comparisons among TOM Estimation Approaches. Focusing on fairly comparing TOM estimation approaches, we fix

[6] http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words/

[7] The provided baselines cover both TREC 2009 and 2010 topics

the TFM estimation approach with MLE. We will observe that only Mix, which follows the requirements for TOM estimation discussed in Section 4.1, obtains consistent improvements over all standard baselines. Specifically, seen from Table 1, we observe that:

1. Gen-MLE improves very slightly or even deteriorates performance over these baselines. The major reason is that general opinion words are not effectively indicative of opinions relevant to the topic. Furthermore, the involvement of topic-unrelated opinion words may cause severe topic–drift when re-ranking the baselines, which largely decreases the topic relevance performance, and consequently, decreases the overall performance.
2. We also note that PRF-MLE shows very remarkable improvements over the two relatively weaker baselines (i.e., stdbaseline2 and stdbaseline3). We argue that this could be mainly attributed to the topic relevance improvements which usually increase the overall performance as well due to the opinionated nature of blogosphere [23]. Indeed, the TOM learned using PRF would highly overlap with factual topic-related words, which help improve topic relevance performance. However, it is infeasible to improve the overall performance over strong topic relevance baselines through only improving topic relevance, since there is very small room for improving topic relevance over these baselines. Thus, we observe slight performance decrease over the strongest baseline (i.e. stdbaseline1) for PRF-MLE. Furthermore, it's naturally more meaningful to improve over strong baselines. Thus, PRF, which could not improve strong baselines, is not a good choice for TOM estimation.
3. In comparison with Mix-MLE and PRF-MLE, Mix-MLE obtains consistent and remarkable improvements over all baselines. The major reason is that Mix can effectively capture language usage of topic-related opinion expressions, which help identify topic-related opinions. This observation verifies the reasonability of requirements for TOM estimation (see Section 4.1) and demonstrates the effectiveness of the proposed approach in Section 4.

Table 1: Performance comparisons among different TOM estimation approaches.

	MAP	p@10	R-prec	Δ MAP(%)
stdbaseline1	0.2427	0.2900	0.2579	–
Mix-MLE	**0.2684**	**0.2950**	**0.2974**	10.58
Gen-MLE	0.2551	**0.2950**	0.2851	5.12
PRF-MLE	0.2409	0.2650	0.2549	-0.75
stdbaseline2	0.1318	0.1700	0.1512	–
Mix-MLE	**0.1531**	**0.2400**	0.1734	16.18
Gen-MLE	0.1305	0.1900	0.1570	-0.98
RPF-MLE	0.1458	0.2050	**0.1837**	10.64
stdbaseline3	0.1001	0.1700	0.1281	–
Mix-MLE	0.1115	0.1800	0.1511	11.40
Gen-MLE	0.1042	0.1750	0.1470	4.14
PRF-MLE	**0.1442**	**0.1950**	**0.1650**	25.35

Table 2 present example results of TOM estimated using Mix and PRF for two TREC topics. Top 20 words with highest probabilities are showed in the table. We can clearly see that Mix can better capture language usage of topic-specific opinion expressions compared with PRF. For instance, for TREC topic 1111 "jazz music", most top words of the trained TOM using Mix are highly topic-specific opinion-bearing words, such as "rhythmic", "melodic", "dreamy", and "superb", while top words of TOM obtained by PRF are highly mixed with factual topic-related words.

Table2: Example results of TOM estimated using Mix and PRF for two TREC topics.

Topic 1111(Jazz music)		Topic 1162(Uzbekistan)	
Mix	PRF	Mix	PRF
musical	jazz	inconclusive	uzbekistan
groove	musicians	unsuitable	uzbek
rhythmic	musical	acne	karimov
mastering	music	foreigner	tashkent
melodic	classical	wealthy	regime
replica	bebop	hubris	fco
unbeatable	improvisation	infest	islamic
swing	chord	intelligible	murray
kindness	compositions	scabies	islam
dreamy	orchestra	unlimited	torture
vocal	composer	dictator	extremism
eclectic	composition	peacefully	asia
indie	genres	guardian	allies
nonesuch	sound	ambiguous	terror
learning	listened	tyranny	central
laughter	saxophone	oppose	western
strenuous	coltrane	unrest	democracy
superb	pop	whispering	samarkand
fiction	listening	servitude	muslims
comedy	soul	terror	foreign

Comparisons among TFM Estimation Approaches. To focus on fairly comparisons for TFM estimation approaches, we fix the TOM estimation approach with GEN. We will observe that TRW, which follows the requirements for TFM estimation discussed in Section 5.1, obtains consistent improvements over all baselines, and in general outperforms other approaches not fully following the requirements. Specifically, seen from Table 3, we observe that:

1. Gen-MLE improves very slightly or even deteriorates performance over three baselines, and Gen-RW outperforms Gen-MLE in terms of most merits over all baselines. This indicates that by capturing salient content in the feed using random walk, which fully exploits both content and structural information in the feed, we could better determine whether the feed has a principal opinionated inclination. This observation verifies reasonability of the first aspect of the requirements.
2. We also note that Gen-TRW further outperforms Gen-RW, and shows consistent improvements in almost all merits over all baselines. This indicates that further considering "topic bias" helps capture topic-related opinions in the feed to better determine whether the feed has a clear on-topic opinionated inclination. This observation verifies reasonability of the second aspect of requirements.

Overall Comparisons. From Table 4, we can observe that Mix-TRW can achieve consistent and remarkable improvements over all standard baselines. We also observe that Mix-TRW consistently outperforms Gen-TRW over all standard baselines in terms of all metrics. Likewise, Mix-TRW outperforms Mix-MLE except for p@10 over stdbaseline3. These observations show the effectiveness and flexibility of our proposed languages modeling approach to integrate both the language usage information of topic-specific opinion expressions and various evidences from the feed to determine whether the feed shows a principal on-topic opinionated inclination.

Table1 3: Performance comparisons among different TFM estimation approaches.

	MAP	p@10	R-prec	Δ MAP(%)
stdbaseline1	0.2427	0.2900	0.2579	_
Gen-TRW	**0.2720**	**0.3050**	**0.2978**	12.09
Gen-MLE	0.2551	0.2950	0.2851	5.12
Gen-RW	0.2630	0.3000	0.2877	8.36
stdbaseline2	0.1318	0.1700	0.1512	_
Gen-TRW	**0.1399**	**0.2150**	**0.1614**	6.16
Gen-MLE	0.1305	0.1900	0.1570	-0.98
Gen-RW	0.1345	0.1950	0.1597	2.00
stdbaseline3	0.1001	0.1700	0.1281	_
Gen-TRW	**0.1151**	0.1700	**0.1470**	15.01
Gen-MLE	0.1042	**0.1750**	**0.1470**	4.14
Gen-RW	0.1148	0.1700	0.1458	14.75

Table4: Performance comparisons among different approaches. Paired t-tests are performed, significant improvements over the corresponding baseline (p-value < 0.05) are marked with *.

	MAP	p@10	R-prec	Δ MAP(%)
stdbaseline1	0.2427	0.2900	0.2579	_
Mix-TRW	**0.2855***	**0.3100***	**0.3036***	**17.61**
Mix-MLE	0.2684	0.2950	0.2974*	10.58
Gen-TRW	0.2720*	0.3050*	0.2978*	12.09
stdbaseline2	0.1318	0.1700	0.1512	_
Mix-TRW	**0.1710***	**0.2500***	**0.1917***	**29.74**
Mix-MLE	0.1531	0.2400*	0.1734*	16.18
Gen-TRW	0.1399	0.2150*	0.1614	6.16
stdbaseline3	0.1001	0.1700	0.1281	_
Mix-TRW	**0.1197***	0.1750	**0.1745***	**19.64**
Mix-MLE	0.1115*	**0.1800**	0.1511*	11.40
Gen-TRW	0.1151	0.1700	0.1470	15.01

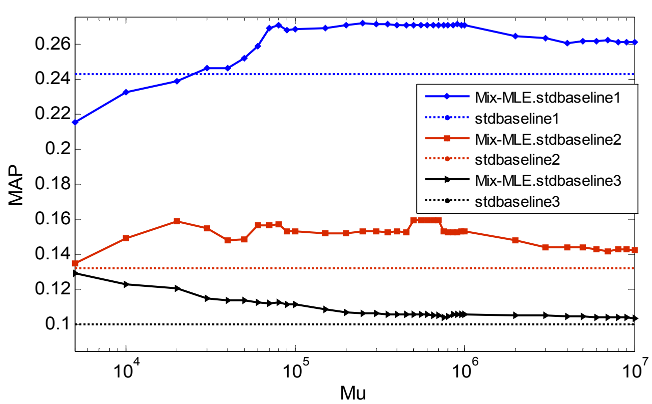

Figure 2: MAP curves with different μ based on standard baselines

Impact of Parameter μ for Estimating TOM with Mix. We here investigate the impact of parameter μ, which controls the influence of the prior for estimating TOM in the mixture model (see Section 4, Equation (6)). We fix the TFM estimation approach with MLE to focusing on purely investigating the impact of μ. We plot the MAP curves with different μ values based on three standard baselines in Figure 2, and from it we can observe that:

1. When μ values are around 100,000, Mix-MLE can obtain remarkable and consistent performance improvements over all standard baselines. And Mix-MLE still have consistent improvements over all baselines within a very large range of μ values (given μ >=60,000). It seems that it is not very sensitive to μ values within this range, especially for strong baselines. This shows the stability and robustness of our proposed mixture model to estimating TOM.
2. When μ values are extremely large (e.g. μ >=1,000,000), the performance go downwards but very smoothly. The major reason is that the learned TOM will be overwhelmed by general opinion words, and thus can not effectively capture topic-specific opinion words.
3. It seems that weaker baselines benefit more from low μ values. Indeed, when μ is very low (e.g., 5,000 and 10,000), Mix-MLE obtains very remarkable performance improvements over the weakest baseline (i.e., stdbaseline3). On the other hand, we observe a remarkable performance decrease over the strongest baseline (i.e., stdbaseline1). The major reason is that, with low μ values, the prior guidance is not adequate to effectively discriminate opinions from factual content. Thus, the learned TOM would highly overlap with factual topic-related words, which helps improve topic relevance performance and, consequently, overall performance for weak baselines, but not for strong baselines.

6.3 Comparisons with TREC approaches

TREC 2009. In TREC 2009, almost all submitted runs deteriorated the performance compared with underlying topic relevance baseline rankings. In fact, Mix-TRW based on stdbaseline1 outperforms the best run (i.e. *ICTNETBDRUN2*) by a large marine (0.2681 vs. 0.1259 in MAP).

TREC 2010. In TREC 2010, three standard baselines were provided by the organizers for fair comparisons of purely opinion-based re-ranking techniques. Mix-TRW based on stdbaseline1 largely outperforms the best run, i.e. *PKUTM111onB1*, among all runs (including those based on standard baselines and not) on 7 TREC 2010 topics (0.3177 vs. 0.2807 in MAP). Besides, most runs based on the standard baselines still provided deteriorated performance over the underlying standard baselines, although some systems managed to improve remarkably over their own baselines. And there was no participant managing to provide consistent and remarkable performance improvements over all standard baselines as our approaches. Table 5 gives the comparisons of Mix-TRW with the best runs based on three standard baselines, respectively. Note that, information in Table5 is based on appendix of the TREC 2010 Proceedings page[8]. Note that, only results on 5 TREC 2010 topics are available in the appendix, the results are based on these 5 topics. Seen from the table, Mix-TRW largely outperforms best runs based on two relatively strong baselines (i.e. stdbaseline1 and stdbaseline2), respectively. And Mix-TRW is defeated by the best run based on the weakest baseline (i.e. stdbaseline3). However, it's naturally more meaningful to improve over stronger baselines.

[8] http://trec.nist.gov/pubs/trec19/t19.proceedings.html

Table5: Performance comparisons with best runs based on standard baselines on TREC 2010 Topics

	MAP
stdbaseline1	0.2128
BIT10std1fd2	0.2240
Mix-TRW	**0.2915**
stdbaseline2	0.1179
ICTNETFBDs2	0.1372
Mix-TRW	**0.2301**
stdbaseline3	0.0927
uogTrfC919s3	**0.1233**
Mix-TRW	0.1086

7. CONCLUSIONS

In this paper, we study opinionated blog feed retrieval, and discuss the challenges of this task. To address these challenges, we propose a unified framework. In this the framework, we propose a language modeling approach to estimating opinion scores, where two language models, Topic-specific Opinion Model (TOM) and Topic-biased Feed Model (TFM), work collectively to reflect whether the blog feed shows a principal on-topic opinionated inclination. We discuss the requirements for an appropriate estimation of TOM and TFM, respectively. Following these requirements, we propose to use a mixture model with prior guidance to estimate TOM and a topic-biased random walk to estimate TFM. In our experiments, we show the reasonability of proposed approaches to estimating TOM and TFM. The experiments also show the effectiveness and flexibility of our proposed languages modeling approach to integrate both the language usage information of topic-specific opinion expressions and various evidences from the feed to improve the performance. As a preliminary work, there may be other ways to estimate the involved two language models better, which will be the focus of our future work.

Acknowledgments

This work was supported by following funds: National Natural Science Foundation of China under Grants, No. 60933005, No. 60873245, No. 60903139, No. 60803085, No.61100083, and No. 60873243; National High-Tech R&D Program of China (863 Program) with Grant No. 2010AA012502, and No. 2010AA012503; National Key Technology R&D program with Grant No. 2011BAH11B02.

8. REFERENCES

[1] Balog, K., Rijke, M., and Weerkamp, W. 2008.Bloggers as Experts Feed Distillation using Expert Retrieval Models. In *Proceedings of SIGIR 2008*.

[2] Elsas, J.L. , Arguello, J., Callan, J., and Carbonell, G.J. 2008. Retrieval and feedback models for blog feed search. In *Proceedings of SIGIR '08*.

[3] Esuli, A., and Sebastiani, F.2005.Determining the semantic orientation of terms through gloss classification. In *Proceedings of CIKM 2005*.

[4] Gerani, S., Carman, M.j., and Crestani, F. 2010. Proximity-based opinion retrieval. In *Proceeding of SIGIR '10.*.

[5] He, B., Macdonald, C., He, J., and Ounis, I. 2008. An effective statistical approach to blog post opinion retrieval. In *Proceeding of CIKM '08*.

[6] Huang, X., and Croft, W. B. 2009. A unified relevance model for opinion retrieval. In *Proceedings of CIKM 2009*, pages 947–956.

[7] Jia, L., Yu, C.2010. UIC at TREC 2010 Faceted Blog Distillation.In *Proceedings of TREC 2010*.

[8] Jiang, P., Zhang, C., Yang, Q., and Niu, Z. 2010. Blog Opinion Retrieval Based on Topic-Opinion Mixture Model. In *Proceedings of PAKDD 2010*.

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

[10] Li, S., Gao, H., Sun, H., Chen, F., Feng, O., Gao, S., Zhang, H., Li, X., Tan, C., Xu, W., Chen, G., and Guo, J. 2009. A Study of Faceted Blog Distillation-- PRIS at TREC 2009 Blog Track. In *Proceedings of TREC 2009*.

[11] Macdonald, C., Ounis, I., and Soboroff, I. 2007. Overview of the TREC 2007 Blog track. In *Proceedings of TREC 2007*.

[12] Macdonald, C., and Ounis, I. 2008. Key blog distillation: ranking aggregates. In *Proceedings CIKM 2008*.

[13] Macdonald, C., Ounis, I., and Soboroff, I. 2009. Overview of the TREC-2009 Blog Track. In *Proceedings of TREC 2009*.

[14] Macdonald, C., Santos, R.L.T., Ounis, I., and Soboroff, I. 2010. Blog track research at TREC. SIGIR Forum 44, 58-75.

[15] McCreadie, R., Macdonald, C., Ounis, I., Peng, J., and Santos, R. 2009. University of Glasgow at TREC 2009: Experiments with Terrier. In *Proceedings of TREC 2009*.

[16] Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. 2007.Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of WWW '07*, 171-180.

[17] Na, S.-H., Lee, Y. , Nam, S.-H., and Lee, J.-H. 2009. Improving opinion retrieval based on query-specific sentiment lexicon. In *Proceedings of ECIR 2009*, pages 734–738.

[18] Ounis, I., Macdonald, C., Rijke, M.de., Mishne, G., and Soboroff, I. 2006.Overview of the TREC 2006 Blog track. In *Proceedings of the 15th Text REtrieval Conference*.

[19] Ounis, I., Macdonald, C., and Soboroff, I.2008.Overview of the TREC-2008 Blog Track. In *Proceedings of TREC'08*.

[20] Ounis, I., Macdonald, C., and Soboroff, I. 2010. Overview of the TREC-2010 Blog Track (Preliminary). In *Proceedings of TREC 2010*.

[21] Sanderson, J.2008. The Blog is Serving Its Purpose: Self-Presentation Strategies on 38pitches.com. *Journal of Computer-Mediated Communication*. Jul 2008, Vol. 13, No. 4: 912-936.

[22] Santos, R. L. T. , He, B., Macdonald, C., and Ounis, I. 2009.Integrating proximity to subjective sentences for blog opinion retrieval. In *Proceedings of ECIR 2009*, pages 325–336.

[23] Seki, K., and Uehara, K. 2009. Adaptive subjective triggers for opinionated document retrieval. In *Proceedings WSDM '09*, 25-33.

[24] Song, L., Cheng, X., Guo, Y., Liu, L., and Ding, G. 2009. ContentEx: A Framework for Automatic Content Extraction Programs. In *Proceedings of ISI'2009*.

[25] Vechtomova, O. 2010.Facet-based opinion retrieval from blogs. *Information Processing and Management*, 46(1):71–88.

[26] Wei, F., Li, W., Lu, Q., and He, Y. 2008. Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In *Proceedings of SIGIR '08*.

[27] Xu, X., Meng, T., Cheng, X., and Liu, Y. 2011. A probabilistic model for opinionated blog feed retrieval. In *Proceedings of the 20th international conference companion on World wide web* (WWW '11).

[28] Zhai C., and Lafferty, J. 2001.Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of CIKM 2001*.

[29] Zhai, C., and Lafferty, J. 2004.A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, Vol. 22, No. 2, 179-214.

[30] Zhai, C., Velivelli, A., and Yu, B.2004. A Cross-Collection Mixture Model for Comparative Text Mining. In *Proceedings of KDD 2004*.

[31] Zhang, W., Yu, C., and Meng, W. 2007. Opinion retrieval from blogs. In *Proceedings of CIKM 2007*, pages 831–840.

[32] Zhang, M., and Ye, X. 2008. A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In *Proceedings of SIGIR 2008*, pages 411–418.

[33] Zhou, Z., Zhang, X.,Vines, P.2010. RMIT at TREC 2010 Blog Track: Faceted Blog Distillation Task. *Online Proceedings of TREC 2010*.