Blog Opinion Retrieval Based on Topic-Opinion Mixture Model

Peng Jiang[1], Chunxia Zhang[2], Qing Yang[1], and Zhendong Niu[1]

[1] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[2] School of Software, Beijing Institute of Technology, Beijing 100081, China
{jp,cxzhang,yangqing2005,zniu}@bit.edu.cn

Abstract. Recently, as blog is becoming a popular medium to express opinions, blog opinion retrieval excites interest in the field of information retrieval. It helps to find and rank blogs by both topic relevance and opinion relevance. This paper presents our topic-opinion mixture model based approach to blog opinion retrieval in the TREC 2009 blog retrieval task. In our approach, we assume each topic has its own opinion relevance model. A topic-opinion mixture model is introduced to update original query model, and can be regarded as a mixture of topic relevance model and opinion relevance model. By pseudo-relevance feedback method, we can estimate these two models from topic relevance feedback documents and opinion relevance feedback documents respectively. Therefore our approach does not need any annotated data to train. In addition, the global representation model is used to represent an entire blog that contains a number of blog posts. Experimental results on TREC blogs08 collection show the effectiveness of our proposed approach.

Keywords: topic-opinion mixture model, blog, opinion retrieval, rank.

1 Introduction

In resent years, blog is becoming an increasingly popular form of communication on the World Wide Web. The blogosphere is a rich information source of public voice, and is useful in extracting and mining public opinions towards some objects or events. Different from other kinds of online textual information, the main characteristics[1] of a blog are: 1) Information provided is often opinion-oriented; 2) Containing numbers of documents that cover a wide range of topics. The need to find appropriate retrieval techniques to track the way bloggers react to products, persons and events raises some challenging problems in the field of information retrieval[2]. Blog opinion retrieval is a task to save the challenge and serve the growing interest in IR.

In this paper, blog opinion retrieval is defined as a task to search blogs with a recurring interest and opinion towards a given topic. Similar to traditional retrieval system, blog opinion retrieval has two basic tasks: 1) search the relevant documents to a user's query, and 2) ranking these documents according to the level of relevance. However, blog opinion retrieval has several special characteristics to be taken into

M.J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, pp. 249–260, 2010.

consideration. The goal of blog opinion retrieval is to find blogs that are principally devoted to certain topics over the time span of the blogs, and to recommend user to subscribe as an interesting feed about the topic (i.e. users may add the interesting feed to their RSS readers). This requires the retrieval unit to be the entire blog containing a number of posts, but not a single post document. Since a blog contains both relevant posts and non-relevant posts to a topic, the overall relevance of a blog must be measured in a proper way. Besides, the blog opinion retrieval goes beyond topic relevance and integrates the opinion relevance in the evaluation of the retrieved blogs. This requires the system to determine whether a blog expresses opinions or facts.

TREC 2009 Blog Track[1] highlights its interest in blog retrieval, and introduces the Faceted Blog Distillation Task. This task takes into account a number of attributes of facets such as opinion, personality and in-depth facets. This paper mainly focuses on the blog retrieval on opinion facet. Technically, there are two typical approaches to the blog opinion retrieval in previous works: two-stage approach based on classification and mixture of language models approach. The two-stage approach is often used in previous TREC Blog Track. There are two basic components in this approach[3]: the retrieval component and the opinion classification component. The former carries out basic relevance retrieval for each query whereas the latter classifies each blog into two categories, namely, opinionated category and factual category. SVM and the maximum entropy classifiers are used in many cases. Mixture of language models approach[4, 5] assumes that a blog is generated by sampling words from a mixture model involving a background language model, a topic language model, and an opinion language model.

In this paper, we present our approach based on the topic-opinion mixture model. It is similar to the above mentioned mixture of language model approach. However, their approaches assume the content of opinion model is the same for all topics, or require models to be trained for every topic by annotated data, or manually input subjective keywords. In our approach, we assume the text opinion expression is dependent on the topic. We first make use of pseudo feedback documents from wiki corpus to construct the topic relevance model, and then some words are automatically selected from a subjective/objective lexicon by the semantic association extent with the topic. Then we combine these words with original query to re-retrieve and get the opinion feedback documents. An opinion relevance model is constructed by these feedback documents. Finally, a topic-opinion mixture model is combined from topic relevance model and opinion relevance model. This model contains topic features and their associated opinion features. So it is effective to evaluate the level of topic relevance and opinion relevance of a blog.

We conduct experiments in this paper on TREC blogs08 datasets, with each blog post being considered as a web page. Moreover, the opinion lexicon (subjective or objective lexicon) used is domain-independent. Hence our proposed approach is applicable to all opinion retrieval tasks on any text resource contained information about topic and opinion, such as product reviews.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the related works in the field. The problem is defined in Section 3. The whole approach is described in Section 4. The experiments and result analysis are presented in Section 5. Finally we conclude the paper and discuss the future work in Section 6.

---

[1] `http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG`

2 Related Works

There are many related works in the TREC Blog Track. First introduced in TREC 2006, the blog track explores the information seeking behavior in the blogosphere. In the past years, the track had two main tasks: the opinion finding task and the blog distillation task. Normally a two-stage process is used to address the opinion finding task. At the first stage, documents are ranked using modern and effective document ranking functions such as BM25[6], language models (LM) and divergence from randomness (DFR) models[7]. A relevance score is allocated to each document. At the second stage of the retrieval process, the classifier [8-12] is used to determine whether a document is opinionated or factual, and an opinionated score is assigned for the document. Next the retrieved documents are re-ranked according to the combined score of the relevance score and the opinion score. Most solutions use a linear combination of relevance score and opinion score, whereas a quadratic combination solution[13] is proposed and achieve a significant improvement.

For the blog distillation task, there are three main solutions: expert finding, pseudo-cluster selection and federated search model. Expert finding solution[7, 14] regards the blog distillation task as an association finding task, between topics and bloggers. blogger model and posting model are proposed for modeling blog distillation[15]. The blogger model represents the blog as a as a multinomial probability distribution over the vocabulary terms. It then computes probability of a query given a blogger. While in the posting model, each post is computed by query likelihood scoring method followed by combining the score for each post. Pseudo-cluster selection solution[16] samples K relevant posts from a blog, and then virtually combines these posts into a topic-dependent pseudo-cluster. Federated search model solution[17] ranks blogs by the estimated number of relevant documents. Pseudo-cluster selection and federated search model solutions use small document model which treats posts of a blog individually. In expert finding solution, large document model which treats all posts of a blog as a whole can achieve a better performance than the small document model. All solutions use language model as the basic retrieval method.

In TREC 2009 Blog Track, the opinion finding task and the blog distillation task are merged into a new task, called faceted blog distillation. Opinion is one of three facets. This paper mainly focuses on the opinion facet. We use a mixture of topic and opinion language models to solve the problem of blog opinion retrieval. A mixture of language models is commonly used in IR application. The basic idea[18] is to infer language models corresponding to unobserved features in the corpus, with the hope that the features learned represent topic and opinion. An example of these works is from Koji and Victor[5], in which sentiment relevance models and topic relevance models are combined based on Generative Models. Mei ant others[4] first introduced Topic-sentiment Mixture model (TSM), which can reveal the latent topical facets in a blog collection, the subtopics in the results of an ad hoc query, and their associated opinions. Their TSM model is a special case of CPLSA model[19], which mixes themes with different views. TSM attempts to learn a general opinion model to all topics, based on the assumption that the opinion model is independent to the topic model. However, in reality, there is a correlation between opinion model and topic model. For example, in topic “wii exercise”, the words represent opinion such as “magical”, “disgust”, “silly” have a higher probability of occurrence; while in topic "westerns movies and novels", the opinionated words such as "flawless", "oddities", "propitiously" are more likely to appear. Our approach assumes each topic has its own opinion relevance model. The opinion relevance model can be estimated by pseudo-relevance feedback, and then combined with topic relevance model which is estimated by wiki pseudo-relevance feedback.

3 Problem Definition

The aim of opinion blog retrieval task is to "find opinionated or factual blogs that are principally devoted to a given topic[2] over the timespan of the blog". Inspired by TREC 2009 Blog Track, we define the opinion blog retrieval task as follows:

Given a topic *T*, find blogs related to *T*, rank them by topic relevance and opinion relevance. The system should provide three blog ranking results according to opinionated relevance, factual relevance and topic relevance as the baseline respectively. The retrieval unit is a blog containing a number of blog posts which can be viewed as web documents.

The previous solution to blog opinion retrieval problem adopted a two-stage strategy: 1) Topic relevance retrieval that finds all topic relevant blogs, regardless of the opinion relevance; 2) Using different classification techniques to compute the opinion relevance of all retrieved blogs, followed by re-ranking them. In the following section, we introduce our approach based on the topic-opinion mixture model to address the blog opinion retrieval task.

4 Our Approach to Blog Opinion Retrieval

4.1 Blog Representation and Query Generation

Following the works of [17, 20], we choose Global Representation Model to represent blog. This model treats a blog as a virtual document which is composed of all posts of the blog. Because this model considers all posts over the timespan of the blog, it can factually reflect the recurring interest of the blog. In addition, since we use language model based approach to rank, Global Representation Model, which combines many posts into a large document, can avoid the problem of sparsity of words as much as possible.

In our approach, title, description and narrative fields of a topic are used for query string generation. First, we filter out unnecessary punctuation marks in the above fields. All verbs are replaced by their infinitives and all nouns by their singular forms. After this, we extract the keywords to build the bag of words. The basic Indri[3] query Q is defined as:

$$\#combine(w_1\ w_2 \ldots w_n)$$

$w_1\ w_2 \ldots w_n$ are the keywords in the bag. We use the following Indri query template to generate query string for a given topic:

[2] Topic in TREC mainly includes three fields: title, description and narrative.

[3] Indri is a search engine from the Lemur project.
http://www.lemurproject.org/indri/

#weight(0.5 Q_{title} 0.3 $Q_{description}$ 0.2 $Q_{narrative}$)

where Q_{title}, $Q_{description}$ and $Q_{narrative}$ are basic Indri queries generated by title, description and narrative field of the topic.

4.2 Basic Retrieval Model

Using the language model approach in IR has shown its effectiveness and simplicity. The general language model approach[21] is decomposed into three components: 1) query model $_Q$; 2) document model $_D$; 3) matching strategy between query model and document model. In our approach, we choose KL-divergence to measure the distance between $_Q$ and $_D$, and rank blogs by the following formula:

$$\begin{aligned} score(D,Q) &= -D(\theta_Q \,\|\, \theta_D) \\ &= -\sum_w p(w \mid \theta_Q) \log \frac{p(w \mid \theta_Q)}{p(w \mid \theta_D)} \\ &= \sum_w p(w \mid \theta_Q) \log p(w \mid \theta_D) + cons(\theta_Q) \end{aligned} \tag{1}$$

Because the constant *cons*($_Q$) does not affect the ranking results, we do not compute it in our system. Thus, the main task is to estimate $_Q$ and $_D$. For blog retrieval in the paper, the document model $_D$ is a multinomial distribution whose parameters are represented by unigram language models. We assume that blog documents are generated by $_D$, which can be estimated by the following formula:

$$p(w \mid \theta_D) = \frac{c(w,D) + \mu p(w \mid C)}{|D| + \mu} = \frac{\sum_{d \in D} c(w,d) + \mu p(w \mid C)}{\sum_{d \in D} |d| + \mu} \tag{2}$$

where $p(w|C)$ is a background language model, d is a post of blog D, $c(w,d)$ is the count of w occurs in d, and μ is a Dirichlet smoothing parameter. We use μ=2000 in this paper, which is optimal in most cases[22].

In traditional approach[21], $_Q$ will be updated by feedback documents model that can be obtained by the relevant documents judged by users, or top documents from initial retrieval. To address the special need for blog opinion retrieval, we introduce Topic-opinion Mixture model $_{TO}$, and interpolate it with the original query model $_Q$ to obtain the updated query model $_{Q'}$, and then assign a score to blog *D* by Formula (1). The updated query model $_{Q'}$ is:

$$\theta_{Q'} = (1-\alpha)\theta_Q + \alpha\theta_{TO} \tag{3}$$

where α controls the influence of topic-opinion mixture model $_{TO}$. In Section 4.3, we describe how to estimate topic-opinion mixture model $_{TO}$.

4.3 Topic-Opinion Mixture Model

The topic-opinion mixture model $_{TO}$ in Formula (3) is the language model which reflects the information need for both topic and opinion; hence a mixture of language models is used to estimate $_{TO}$. In our solution, we define two language models, namely, topic relevance model $_{T}$ and opinion relevance model $_{O}$. The topic-opinion mixture model $_{TO}$ is a linear combination of the two language models:

$$\theta_{TO} = (1-\beta)\theta_T + \beta\theta_O \tag{4}$$

where β is used to control influence of opinion relevance model $_{O}$.

In general, the topic relevance model $_{T}$ in Formula (4) can be obtained by pseudo-relevance feedback method (PRF). PRF assumes the *k* top-retrieved documents are relevant to the original query and extracts highly discriminative words from those documents to update the original query model. We use divergence minimization algorithm[21] to estimate $_{T}$. The divergence minimization algorithm assumes that the topic relevance model is very close to each language model of feedback documents, and uses KL-divergence as the distance between two language models. In order to obtain the feedback documents with high relevance, we index the Wikipedia corpus[4] and treat the *k* top-retrieved wiki pages as the relevance feedback documents. Given a topic *T*, let $F = \{d_1, \ldots d_k\}$ be a set of top *k* retrieved feedback documents from Wikipedia corpus. So the distance can be represented as:

$$D(\theta_T, F) = \frac{1}{k}\sum_{i=1}^{k} D(\theta_T \,\|\, \theta_{d_i}) - \lambda D(\theta_T \,\|\, \theta_{Wiki}) \tag{5}$$

Where $_{Wiki}$ is the Wikipedia corpus language model, $\lambda \in [0, 1)$ is the factor that controls the weight of Wikipedia corpus language model. Following [21], $p(w|_{T})$ can be computed as follows:

$$p(w \mid \theta_T) \propto \exp\left(\frac{1}{1-\lambda}\frac{1}{k}\sum_{i=1}^{k}\log p(w \mid \theta_{d_i}) - \frac{1}{1-\lambda}\log p(w \mid \theta_{Wiki})\right) \tag{6}$$

According to Formula (6), words that are common in the feedback documents, but not common in the entire Wiki corpus will be assigned a higher probability. In our system, k=25, λ=0.5, the feedback terms count is set to be 100.

Next we must estimate the opinion relevance model $_{O}$ in Formula (4). $_{O}$ reflects the users' information need for opinion. Some bloggers provide opinionated content for their interested topics, while others report factual information. So we need to estimate two $_{O}$, one for opinionated information and the other for factual information. Previous works show that the opinion always has an association with topic. Different topics may have a different opinion expression. But training different models on annotated data for different topic is usually unpractical.

The basic procedure of our approach has two steps. The first step is to expand original query with some subjective words or objective words, and then use the expanded query to obtain the top *k* ranked results as pseudo-feedback documents. The second step is to make use of pseudo-relevance feedback method to estimate $_{O}$. For the first step, the most important thing is to select *m* subjective/objective words that have the closest association with a given topic. In our solution, we use a subjective lexicon and an objective lexicon. The subjective lexicon contains 8821 words that are

[4] http://download.wikimedia.org/enwiki/

used in OpinionFinder[23]. The words in objective lexicon are selected from SentiWordNet[24]. Similar to [25], we use the Pointwise Mutual Information (PMI) to measure the semantic association between subjective/objective word w and the query string Q of a given topic:

$$PMI\left(w,Q\right)=\log\frac{p\left(w,Q\right)}{p\left(w\right)p\left(Q\right)}=\log\frac{hits\left(\#uw15\left(w\quad Q\right)\right)\times\left|C\right|}{hits\left(w\right)\times hits\left(Q\right)} \tag{7}$$

where $|C|$ is the total number of documents in corpus. We make use of blog collection index to estimate PMI. $hits(w)$ and $hits(Q)$ are the counts of retrieved documents which contain subjective/objective word w and query string Q respectively. $hits(\#uw15(w\quad Q))$ is the count of retrieved documents containing w and Q simultaneously in an unordered window of 15 terms. The reason why we use a fixed size window instead of a sentence is that: it is time-consuming and unpractical to split all text into sentences, and the inaccuracy can be ignored when large corpus is used. To avoid division by zero, 0.01 is added to the number of hits. Finally we choose the top 30 subjective/objective words according to the PMI value, and use them to expand original query. The feedback documents can be used to build opinion relevance model $_O$ by Formula (6).

5 Experiments

5.1 Experiment Setup

5.1.1 Data Sets

We use TREC Blogs08 collection as required by TREC 2009 Blog Track to evaluate our approach. The summary statistics of this collection is shown in Table 1. We actually use the permalinks and homepages in our approach. Blog feeds collection is not used. It is because the text in the feed pages usually contains a few sentences of each post and therefore cannot reflect the topic or opinion well. The permalinks and homepages are encoded by HTML. We use Indri to index them respectively. The Krovetz stemmer and a list with 450 stop words are used to pre-process.

Table 1. Summary statistics of data sets

Data Set	Doc number	Size (Uncompressed)	Time span
homepages	1,011,733	56G	14/01/2008
feeds	1,303,520	808G	~
permalinks	28,488,767	1445G	10/02/2009

5.1.2 Evaluation

There are 13 opinion topics provided by TREC 2009 Blog Track (see Table 2). The evaluation metrics used are standard IR measures[26], such as mean average precision (MAP), R-Precision (R-prec), and precision at top 10 results (p@10). The relevance and opinion judgments adopt the TREC 2009 Blog Track standards: not judged (-1), not relevant (0), relevant (1), relevant and opinionated (2) and relevant and factual (3). All results are assessed by the evaluation tool provided by TREC.

There are four approaches in our experiments for comparative studies: (1) Our Topic-opinion Mixture Model (TOM) (2) MEClassifier. It is a traditional approach based on classifier. We trained a maximum entropy classifier on Movie Review Data. The classifier takes blog text vector as input, and outputs opinionated or factual label and an associated score, which is combined with original relevance score. Blogs is then re-ranked by the combined score. (3) SingleModel. It combines all topic models with the same opinion model. This approach is introduced in [4], which treats the opinion model the same for all topics in a collection. (4) Baseline. It only considers the topic relevance score while ranking the opinionated and the actual blogs.

Table 2. Opinion topics in TREC Blog 2009

No.	Title	No.	Title	No.	Title
1103	farm subsidies	1125	cosmetic surgery	1141	sciatica remedies
1106	taiwan politics	1132	gun control dc	1144	future of journalism
1111	jazz music	1134	new orleans after katrina	1150	NASA space program
1116	homeopathic medicine	1137	civil unions		
1119	no child left behind	1140	scientology		

5.2 Experimental Results

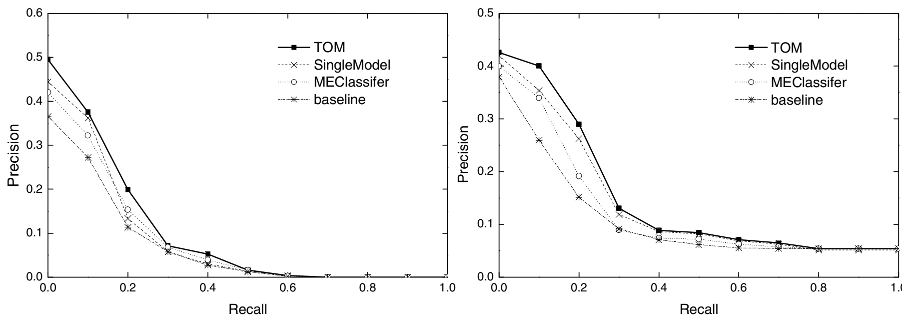
5.2.1 Overview of Experimental Results

Result comparisons of each approach are presented in Table 3 and Fig.1. The results show that all approaches outperform the baseline. Comparing with other approach, our approach achieves the best retrieval performances except for R-prec and P@10 of factual blog retrieval in Table 3. This demonstrates that our proposed approach is effective especially for opinionated blog retrieval.

Fig. 2 (a) and (b) show the performance improvements over baseline on each topic in terms of MAP and R-prec. The average improvements on all topics for opinionated blogs retrieval are 48.87% and 26.39% in terms of MAP and R-prec. The average improvements for factual blogs retrieval are 22.69% and 8.82% in terms of MAP and R-prec. We note that there is a slight improvement over baseline in factual blog retrieval. The explanation is that, ranking by topic and factual relevance does not have much difference from ranking only by topic relevance. Only topic 1134 and 1150 get decreased performance. In terms of MAP, there are 5 topics which have no improvement over baseline for factual blogs retrieval, comparing with 2 topics for opinion blogs retrieval. In terms of R-prec, there are 7 topics which have no improvement over baseline for factual blogs retrieval, comparing with 5 topics for opinion blogs retrieval. This proves that our approach is more effective for opinionated blogs retrieval than factual blogs retrieval.

Table 3. Performance comparison among different approaches

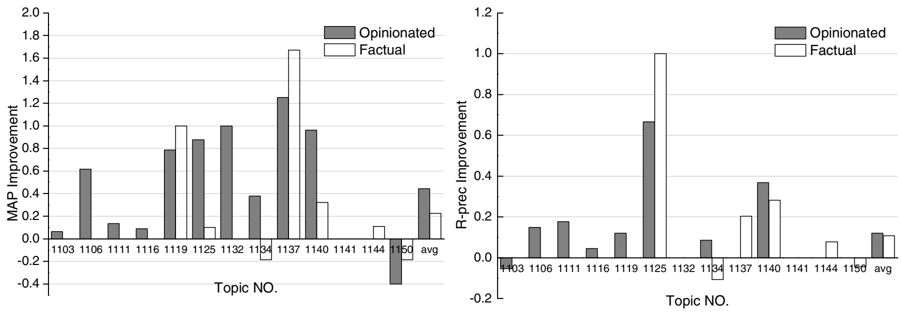
Approaches	MAP		R-prec		P@10	
	opinionated	factual	opinionated	factual	opinionated	factual
Baseline	0.0573	0.1124	0.1027	0.1270	0.0923	0.0846
MEClassifer	0.0693	0.1236	0.1298	**0.1402**	0.1000	0.1077
SingleModel	0.0732	0.1159	0.1302	0.1305	0.1154	**0.1231**
TOM	**0.0853**	**0.1379**	**0.1317**	0.1382	**0.1231**	0.1154



(a) Opinionated blogs retrieval

(b) Factual blogs retrieval

Fig. 1. Comparison of recall-precision curves among different approaches



(a) MAP improvement over baseline

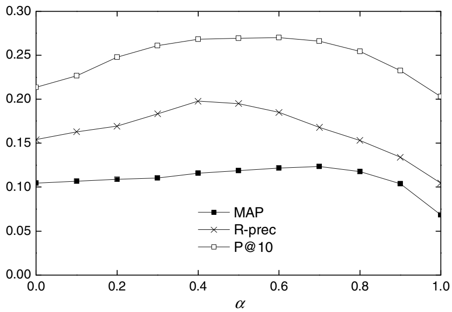
(b) R-prec improvement over baseline

Fig. 2. Performance improvements over baseline on each topic

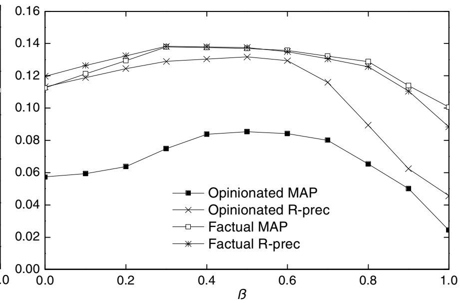
5.2.2 Analysis of Parameters of Topic-Opinion Mixture Model

In our approach, the parameter β of the topic-opinion mixture model controls influence of opinion relevance model $_O$. Specifically, β is used to adjust the ratio of topic relevance and opinion relevance in topic-opinion mixture model. In order to analyze the effect of β, we note that parameter α in Formula (3) may affect the final performance. The difference can be observed in Fig. 3 (a), in which we show the changing performances by changing α from 0 to 1, with a step up size of 0.1. In this experiment, we set β=0, thus, $_{TO}$ actually becomes the topic relevance model $_T$. Therefore the experiment actually evaluates the effects of feedback documents from Wiki corpus. We notice that using feedback model from wiki documents can generally improve the performance. But when it is too large approaching 1, the performance is extremely bad and is even worse than the performance without using feedback model. We choose α=0.5, which is a value that can usually achieve better performance than other values.

Fig. 3 (b) shows how MAP, R-prec varies accordingly with β, when α is fixed at 0.5. Note that performance at β=0 is actually the baseline performance. Overall, when the β value increases, the overall performance improves. But when β is too large, the overall performance deteriorates sharply. Be more specific, when β=0.5 the opinionated blog retrieval achieves its best performance; when β=0.3 the factual blog retrieval achieves its best performance. This is because the topic relevance model helps to focus on the topic, while the opinion relevance model can supplement subjective or objective words for the purpose of opinion retrieval. When β is too large, there will be many opinionated or factual blogs with no topic relevance.



(a) Performance sensitivity to α



(b) Performance sensitivity to β

Fig. 3. Performance sensitivity to parameters

5.2.3 Analysis of Samples from Topic-Opinion Mixture Model

Table 4 presents sample probabilities using topic-opinion mixture model. Samples are divided into the two topics: "jazz music" and "no child left behind". The "Topic model" columns contain the topic words. These words may come from the subtopic of the corresponding topic, such as "musician", "band", "Africa", "educate", "fund", etc. So they can be treated as supplement for the original query. The "Opinionated model" columns contain subjective words related to the corresponding topic. As we have discussed above, the opinionated relevance model varies significantly with topics. For instance, for "jazz music" topic, the subjective words "limitless", "entertaining" have relatively higher probability of occurrence; whereas for "no child left behind" topic, the associated subjective words are "willing", "supportive", etc. In the "Factual model" columns, the words are found to be neutral, without any semantic orientation. Some words appear in many topics, such as "comment", "state", etc. This reflects that the factual relevance model has low association with topics.

Table 4. Sample probabilities from topic-opinion mixture model. The top 10 words with high probability of occurrence are selected. Results of two topics are presented corresponding to the three language models: topic relevance model, opinionated model and factual model.

Topic 1111 jazz music						Topic 1119 no child left behind					
Topic model		Opinionated model		Factual model		Topic model		Opinionated model		Factual model	
w	$p(w\|\theta_T)$	w	$p(w\|\theta_O)$	w	$p(w\|\theta_O)$	w	$p(w\|\theta_T)$	w	$p(w\|\theta_O)$	w	$p(w\|\theta_O)$
jazz	0.0730	exclusive	0.0137	comment	0.0141	school	0.0343	willing	0.0036	comment	0.0194
music	0.0303	like	0.0137	new	0.0119	student	0.0313	rightly	0.0035	learn	0.0183
play	0.0163	inestimably	0.0040	clear	0.0041	state	0.0263	supportive	0.0035	state	0.0062
musician	0.0148	limitless	0.0040	state	0.0040	nclb	0.0258	benefit	0.0035	question	0.0057
style	0.0133	entertaining	0.0026	profile	0.0039	educate	0.0223	clearly	0.0035	address	0.0052
blue	0.0119	goodly	0.0023	concert	0.0038	fund	0.0219	contentment	0.0035	break	0.0048
new	0.0119	friendly	0.0020	old	0.0037	federal	0.0119	important	0.0035	require	0.0047
band	0.0111	willing	0.0017	live	0.0037	assess	0.0104	transparent	0.0034	public	0.0040
america	0.0107	great	0.0015	swing	0.0032	child	0.0074	winnable	0.0034	legal	0.0039
africa	0.0100	creative	0.0013	classic	0.0031	support	0.0070	justly	0.0027	educational	0.0038

6 Conclusions

In this paper, we present an approach to the task of blog opinion retrieval. This approach uses topic-opinion mixture model to solve the problem of ranking blog not only by topic relevance but also by opinion relevance. Comparing with previous work, this model can effectively learn opinion relevance model without training on annotated data. In addition, the opinion relevance models vary with topics so that the model's effectiveness to different topics is ensured. We evaluate our model on TREC Blogs08 collection, and the experimental results show that the topic-opinion mixture model approach achieves a better performance than other approaches for most of the opinion topics in TREC 2009 Blog Track.

In general, performance of the blog opinion retrieval is worse than traditional text retrieval. There is still a huge potential space for further research to improve the performance of blog opinion retrieval. In addition, it would be interesting to explore the knowledge behind topic and opinion from the perspective of time dimension of blogs. Another interesting future research direction is to use the mixture language model to explore the other blog attributes or facets such as writing style, authority, etc.

Acknowledgments. We would like to express our gratitude to Computer Software Lab of Beijing Institute of Technology for providing us experiment condition for our participation in TREC 09. This work is supported by the grant from Chinese National Natural Science Foundation (No: 60705022).

References

1. Marti, A.H., Matthew, H., Susan, T.D.: What should blog search look like? In: Proceeding of the 2008 ACM workshop on Search in social media. ACM, Napa Valley (2008)
2. Ounis, I., Macdonald, C., Soboroff, I.: On the TREC blog track. In: Proceedings of the International Conference on Weblogs and Social Media (ICWSM), Seattle, USA (2008)
3. Liu, B.: Sentiment Analysis and Subjectivity. CRC Press, Taylor and Francis Group (2009)
4. Qiaozhu, M., Xu, L., Matthew, W., Hang, S., ChengXiang, Z.: Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th international conference on World Wide Web, pp. 171–180. ACM, Banff (2007)
5. Koji, E., Victor, L.: Sentiment retrieval using generative models. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 345–354. Association for Computational Linguistics, Sydney (2006)
6. Lee, Y., Na, S.H., Kim, J., Nam, S.H., Jung, H.Y., Lee, J.H.: Kle at trec 2008 blog track: Blog post and feed retrieval. In: Proceedings of TREC-08 (2008)
7. He, B., Macdonald, C., Ounis, I., Peng, J., Santos, R.L.T.: University of glasgow at trec 2008: Experiments in blog, enterprise, and relevance feedback tracks with terrier. In: Proceedings of TREC-08 (2008)
8. Bermingham, A., Smeaton, A., Foster, J., Hogan, D.: DCU at the TREC 2008 Blog Track. In: Proceedings of TREC-08 (2008)
9. Hoang, L., Lee, S.W., Hong, G., Lee, J.Y., Rim, H.C.: A Hybrid Method for Opinion Finding Task (KUNLP at TREC 2008 Blog Track). In: Proceedings of TREC-08 (2008)

10. Jia, L., Yu, C., Zhang, W.: UIC at TREC 2008 blog track. In: Proceedings of TREC-08 (2008)
11. Li, B., Liu, F., Liu, Y.: UTDallas at TREC 2008 Blog Track. In: Proceedings of TREC-08 (2008)
12. He, H., Chen, B., Du, L., Li, S., Gao, H., Xu, W., Guo, J.: PRIS in TREC 2008 Blog Track. In: Proceedings of TREC-08 (2008)
13. Min, Z., Xingyao, Y.: A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 411–418. ACM, Singapore (2008)
14. Weerkamp, W., Rijke, M.d.: External Query Expansion in the Blogosphere. In: Proceedings of TREC-08 (2008)
15. Balog, K., de Rijke, M., Weerkamp, W.: Bloggers as experts. In: 31st Annual International ACM SIGIR Conference (SIGIR 2008), pp. 753–754. ACM, Singapore (2008)
16. Seo, J., Croft, W.B.: UMass at TREC 2008 Blog Distillation Task. In: Proceedings of TREC-08 (2008)
17. Jonathan, L.E., Jaime, A., Jamie, C., Jaime, G.C.: Retrieval and feedback models for blog feed search. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 347–354. ACM, Singapore (2008)
18. Bo, P., Lillian, L.: Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr. 2, 1–135 (2008)
19. Qiaozhu, M., ChengXiang, Z.: A mixture model for contextual text mining. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 649–655. ACM, Philadelphia (2006)
20. Jangwon, S., Croft, W.B.: Blog site search using resource selection. In: Proceeding of the 17th ACM conference on Information and knowledge management, pp. 1053–1062. ACM, Napa Valley (2008)
21. Chengxiang, Z., John, L.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of the tenth international conference on Information and knowledge management. ACM, Atlanta (2001)
22. Chengxiang, Z., John, L.: A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst. 22, 179–214 (2004)
23. Theresa, W., Janyce, W., Paul, H.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 347–354. Association for Computational Linguistics, Vancouver (2005)
24. Esuli, A., Sebastiani, F.: SentiWordNet: A publicly available lexical resource for opinion mining. In: Proceedings of LREC, pp. 417–422 (2006)
25. Turney, P.D., Michael, L.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Trans. Inf. Syst. 21, 315–346 (2003)
26. Chris, B., Ellen, M.V.: Retrieval evaluation with incomplete information. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 25–32. ACM, Sheffield (2004)