A Probabilistic Model for Opinionated Blog Feed Retrieval

Xueke Xu[1,2], Tao Meng[1], Xueqi Cheng[1], Yue Liu[1]
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
2. Graduate School of Chinese Academy of Sciences, Beijing, China
{xuxueke, mengtao}@software.ict.ac.cn, {cxq,liuyue }@ict.ac.cn

ABSTRACT
In this poster, we study the problem of *Opinionated Blog Feed Retrieval* which can be considered as a particular type of the *faceted blog distillation* introduced by TREC 2009. It is a task of finding blogs not only having a principle and recurring interest in a given topic but also having a clear inclination towards expressing opinions on it. We propose a novel probabilistic model for this task which combines its two factors, topical relevance and opinionatedness, in a unified probabilistic framework. Experiments conducted in the context of the TREC 2009 & 2010 Blog Track show the effectiveness of the proposed model.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Retrieval Model
General Terms
Algorithms, Performance, Experimentation.
Keywords
opinionated blog feed retrieval, topical relevance, opinionatedness

1. INTRODUCTION
Opinionated Blog Feed Retrieval is a pilot task that can be defined as identifying blog feeds not only having a principle and recurring interest in the given topic but also having a clear inclination towards expressing opinions on it. Specifically, given a query, the system should provide a list of ranked opinionated relevant blog feeds. Those top ranked will be recommended to users for RSS subscriptions, and via this recommendation, users may track public opinions on their interesting topics in time. This task can be considered as a particular type of the *faceted blog distillation* introduced by TREC 2009[3, 4], and corresponds to the "*opinionated*" value for the "**opinionated**" facet.

Responding to the requirements of the task, two factors should be considered for the blog ranking here: topical relevance and opinionatedness on the topic. Most existing approaches presented in TREC 2009 consider them separately, and perform in a two-stage way: first generate a topical baseline regardless of opinionatedness, and next estimate opinionatedness to re-rank the topical baseline with a heuristic manner. The opinionatedness estimation can be conducted with classification techniques or an opinion lexicon. However, in almost cases, when opinionatedness is considered, a decrease in performance is observed compared to the baseline [3]. This motives us to adopt a unified approach to combine both of the factors for this task.

In this poster, we propose a probabilistic model for this task, and it has the following characteristics:

1. It combines topical relevance and opinionatedness on the topic in a unified probabilistic framework. The combination can be directly interpreted in a principled way, while most existing work uses a heuristic manner.
2. The opinionatedness is estimated topic-dependently, which is embodied in two aspects: (1) first, for each topic, we construct a topic-specific opinion model to represent the topic-specific opinion expressions; (2) second, for each blog, we measure topic-sensitive weight of each word in the blog to reflect topic-biased content characteristics.
3. The model requires no training, and thus does not require manual labeling. Furthermore, it needs no additional external resource, but a general opinion lexicon. These points can minimize human labor.

We conduct experiments in the context of the TREC 2009 & 2010 Blog Track. The experimental results show that our model can remarkably improve the performance over topical baseline.

WWW 2011, March 28-April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

2. THE PROPOSED MODEL
Our model aims to develop an effective function that ranks blogs considering both topical relevance and opinionatedness. According to traditional generative model in Information Retrieval (IR) area, topical relevance can be estimated by its generation likelihood given the query Q, $P(blog|Q)$. Following this generative model, for our task, we further consider opinionatedness. Thus we introduce the latent variable O_Q which indicates the topic-specific opinion expressions, and rank blogs according to their generation probability given the query Q and O_Q, $P(blog|Q,O_Q)$. Formally,

$$P(blog|Q,O_Q) \propto P(blog,Q,O_Q) = P(blog)P(Q|blog)P(O_Q|Q,blog)$$
$$= P(blog)P(Q|blog)\sum_{w\in V} P(w|Q,blog)P(O_Q|w) \quad (1)$$

By assuming the prior probability of each word w (i.e., $P(w)$) to be uniform, we have the following equation:

$$P(O_Q|w) = \frac{P(w|O_Q)P(O_Q)}{P(w)} \propto P(w|O_Q) \quad (2)$$

Plugging Equation (2) into Equation (1), we come to the following equation:

$$P(blog|Q,O_Q) \propto P(blog)P(Q|blog)\sum_{w\in V} P(w|Q,blog)P(w|O_Q) \quad (3)$$

There are two major components in Equation (3). $P(blog)p(Q|blog)$ considers the topical relevance, and $\sum_{w\in V} P(w|Q,blog)P(w|O_Q)$ deals with the opinionatedness.

3. TOPICAL RELEVANCE ESTIMATION
Since $P(blog)p(Q|blog)$ deals with the topical relevance, it can be estimated by the existing approaches to topical blog feed search. In this poster, we adopt the Small Document (SD) model [1]. According to the SD model, each blog is considered as a collection of its constituent posts, and $P(blog)p(Q|blog)$ can be given as follows:

$$P(blog)P(Q|blog) = P(blog)\sum_{post\in blog} P(Q|post)P(post|blog) \quad (4)$$

where $P(blog)$ is the blog prior probability computed as $\log(N_{blog})$ to favor the blogs with more posts, here N_{blog} is the number of posts in the blog; $P(Q|post)$ is the query likelihood of the post, which is computed using the BM25 model in this poster; and $P(post|blog)$ is the post centrality, which we assume to be uniform.

4. OPINIONATEDNESS ESTIMATION

$\sum_{w\in V} P(w|Q,blog)P(w|O_Q)$ estimates the opinionatedness on the topic, where O_Q can be represented as a language model(referred to as O_Q LM), and $P(w|O_Q)$ is the probability of the word w in O_Q LM; $P(w|Q,blog)$ is the probability of w given the blog and the query Q, measuring the topic-sensitive weight of w in the blog.

4.1 Estimating O_Q LM

We beforehand collect a general opinion lexicon[1]. Then, given a query Q, we can learn the O_Q LM as follows: (1) use the original query to retrieve the top N topically relevant posts from the TREC Blogs08 collection with the BM25 model; (2) use all the general opinion words together as a query to re-retrieve the top K posts as opinion feedback documents A from the top N topically relevant posts retrieved in the step (1)(K N, in our experiments, K=30, and N=15000); (3)use the Bol term weighting model [2] to assign a weight to each word in the vocabulary V, measuring how informative it is in A against the background collection(i.e., Blogs08 collection in our experiments), to infer the probability of the word in the LM. The words with high probability in the LM should be topic-related opinion words, or indicate controversial subtopics on which bloggers tend to express opinion.

4.2 Estimating $P(w|Q,blog)$

By considering each blog as a collection of its constituent posts like SD model mentioned in Section 3, we have:

$$P(w|Q,blog) = \sum_{post\in blog} P(post|Q,blog)P(w|post) \quad (5)$$

By approximating $P(post|Q,blog)$ with $P(post|Q)$, we have:

$$P(w|Q,blog) = \sum_{post\in blog} P(post|Q)P(w|post)$$
$$\propto \sum_{post\in blog} P(Q|post)P(w|post) \quad (6)$$

where $P(Q|post)$ measures the topical relevance of the post using the BM25 model, and $P(w|post)$ is estimated using Maximum Likelihood Estimation(MLE) with Dirichlet smoothing.

Finally, for each word w, $P(w|Q,blog)$ is calculated as the sum of its probability in the constituent posts, weighted by the post topical relevance. Compared to $P(w|blog)$, $P(w|Q,blog)$ assigns more probability to the words more related to the topic to reflect topic-biased content characteristics of the blog. It may serve as the weighting factor for the aggregation of opinion expressions (i.e., $P(w|O_Q)$) within the blog to highlight those really towards the topic.

5. EXPERIMENTS & RESULTS

We conduct experiments in context of the *faceted blog distillation* task of the TREC 2009 & 2010 Blog Track, and use the TREC Blogs08 collection. In the *faceted blog distillation* task, each query is associated with an additional "*facet*" field besides the traditional query fields. Each facet has two values, and each value corresponds to a ranking of blogs respectively. Our task only considers the "*opinionated*" value for the "**opinionated**" facet. There are totally 20 "**opinionated**" topics in TREC 2009&2010 Blog Track officially used for evaluation of this task (including 13 topics for 2009 and 7 topics for 2010). We use all these topics.

Table 1: Performance comparisons among different approaches

	2009 topics			2010 topics		
	MAP	R-prec	p@10	MAP	R-prec	p@10
Unified Model	**.2434**	**.2469**	**.2615**	**.1500**	**.1894**	**.2857**
Topical Baseline	.1655	.1752	.1923	.1173	.1830	.1857
Unified Model O	.2008	.2246	.2461	.1245	.1700	.2286

In Table 1, *Topical Baseline* is the topical relevance component of our model(i.e., $P(blog)p(Q|blog)$); *Unified Model O* uses the general opinion expressions *O,* which is assumed to be the general opinion lexicon with each opinion word uniformly distributed, instead of O_Q. Table 1 shows that the MAP improvements of our *unified model* over the *Topical Baseline* on 2009 and 2010 topics are 47.07% and 27.9% respectively. Besides, our *unified model* outperforms the best run for this task in TREC 2009 Blog Track, whose MAP value is 0.1295 on 2009 topics, by a large margin. The table also shows the performance benefits greatly from using O_Q compared to using O, which verifies the reasonability of modeling topic-specific opinion expressions for this task.

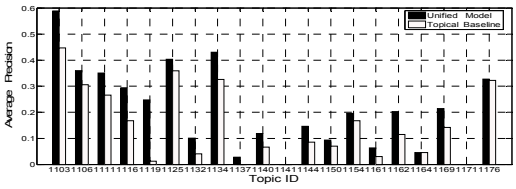

Figure 1: Performance comparisons on each topic

Figure 1 shows the performance comparisons with the *Topical Baseline* on each topic. We can observe the improvements over *Topical Baseline* on most topics (18 of 20) and no performance decrease on any topic, which indicates the stability of our model.

6. FUTURE WORK

In this poster, we propose a probabilistic model for Opinionated Blog Feed Retrieval. For future work, we plan to try more reasonable approach to estimating O_Q LM to better capture the opinions relevant to the topic.

7. ACKNOWLEDGMENTS

This work was mainly funded by National Natural Science Foundation of China under grant number 60873245, 60903139.

8. REFERENCES

[1]Elsas, J., Arguello, J., Callan, J., and Carbonell, J. 2008. Retrieval and feedback models for blog feed search, In *Proceedings SIGIR 2008.*

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

[3]Macdonald, C., Ounis, I., and Soboroff, I. 2010. Overview of the TREC-2009 Blog Track. In *Proceedings of TREC 2009.*

[4]TREC Blog track wiki. http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG, 2010.

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