Efficient Feedback-based Feature Learning for Blog Distillation as a Terabyte Challenge

Dehong Gao, Wenjie Li, Renxian Zhang

Department of Computing, the Hong Kong Polytechnic University. Hong Kong

{csdgao, cswjli, csrzhang}@comp.polyu.edu.hk

ABSTRACT

The paper is focused on blogosphere research based on the TREC blog distillation task, and aims to explore unbiased and significant features automatically and efficiently. Feedback from faceted feeds is introduced to harvest relevant features and information gain is used to select discriminative features, including the unigrams as well as the patterns of unigram associations. Meanwhile facing the terabyte blog dataset, some flexible processing is adopted in our approach. The evaluation result shows that the selected feedback features can greatly improve the performance and adapt well to the terabyte data.

KEYWORDS : Blog Distillation, Faceted Distillation, Feedback

1. Introduction

With the accelerated growth of social networks, both organizations and individuals have shown great interest in conveying or exchanging ideas and opinions. The blogosphere provides an ideal platform for communication. According to the statistics of Blogpulse(*blogpulse.com*) in Jan. 2011, more than 152 million blogs have been published. One interesting issue related to the massive blogs is to automatically explore authors' behaviours from their blog posts.

Research related to blog posts mainly focuses on opinion retrieval to identify topic-based opinion posts, which means retrieved posts should not only be relevant to the targets, but also contain subjective opinions about given topics. Generally, topic-based opinion retrieval can be divided into two parts using a separated relevance and opinion model or a unified relevance model. In separated models, posts are first ranked by topical relevance only, then, the opinion scores can be acquired by either classifiers, such as SVM (Wu Zhang and Clement Yu, 2007), or external toolkits like OpinionFinder(*www.cs.pitt.edu/mpqa/*) (David Hannah et al. 2007). The precision of the opinion retrieval is highly dependent on the precision of relevance retrieval. Huang et al propose a unified relevance model by integrating a query-dependent and a query-independent method, which achieved high performance in topic-based opinion retrieval (Huang et al., 2009).

Based on opinionated blogosphere identification, TREC introduces the faceted blog distillation track in 2009 with two subtasks: baseline distillation and faceted distillation. The former is to retrieve all the relevant feeds corresponding to given topics without any consideration of facets. The latter aims to re-rank the baseline feeds according to specific facets. For operational simplicity, TREC specifies three faceted inclination pairs (TREC Blog Wiki, 2009):

Opinionated vs. Factual inclinations aim to differentiate feeds expressing opinions from those describing factual information;

Personal vs. Official inclinations are to discriminate individual-authored feeds from organization-issued ones;

Proceedings of COLING 2012: Demonstration Papers, pages 147–154,
COLING 2012, Mumbai, December 2012.

In-depth vs. shallow inclinations have the purpose of separating feeds involving deep analysis from those conveying shallow thoughts.

So far, several methods have been attempted for faceted distillation. In (Richard mcCreadie et al, 2009), SVM and ME classifiers are introduced to predict the faceted inclinations of each feed according to pre-trained models. In (Mostafa Keikaha et al., 2009), feed faceted scores are heuristically given to re-rank feeds. For classification as well as re-ranking, the challenge is to select the features related to each inclination. Most work at present focuses on exploring heuristic features. For example, length of the posts is introduced for in-depth and shallow inclination ranking and occurrence of personal pronouns also serves as a feature for personal and factual inclination ranking (Mejova Yelena et al., 2009). Other heuristic features, like permalink number and comment number, are also commonly used in these inclination ranking (SeungHoon Na et al., 2009; Bouma Gerlof 2009; Kiyomi Chujo et al, 2010). However, we observe that for some facets these features are far from enough. For example, it is really hard to discover the indicative heuristic features for some facets like factual, personal and official inclinations. In view of this, we attempt to introduce some terms as common features from blog corpus. Cooperating with faceted feedback information, we first discover more (non-heuristic) feature candidates, including unigram features as well as some word collaborated patterns features, e.g., combination of unigrams "company" and "report", etc. These features are then selected by feature selection, in particular with point-wise mutual information. Furthermore, since the size of our experiment dataset is up to terabyte, we take some flexible processing to adapt to the massive dataset, which has been proved to be efficient in our experiments. In a word, we believe the benefits of this work can be twofold. (1) Rather than only using heuristic features, we can learn more faceted related features automatically, and this method can be directly applied in new defined facets. (2) By some flexible processing, our work is quite efficient for massive dataset.

2. Feedback Feature Learning

Our following work is based on the blog distillation, and as mentioned in the above section, baseline distillation subtask needs to be conducted before feature learning in faceted distillation. Thus, we first briefly introduce the baseline distillation. To enhance efficiency in the face of the huge and noisy raw data (2.3TB), we implement a distributed system and adopt the Indri tool(*www.lemurproject.org*) for our purpose, with its default language model and related toolkits. With the help of these tools, the top related feeds can be retrieved according to given topics in the baseline distillation.

Based on the ranking of the baseline distillation, we then focus on the **faceted blog distillation** and **feedback feature learning**. The key issue in faceted blog distillation is to discover the relevant and discriminative features for each faceted inclination, and determine the weight of each feature. To solve the issue above, our approach explores features from three orientations: *Heuristic Features* (*HF*), *available Lexicon Features* (*LF*), and *Corpora Learned Features* (*CF*).

Heuristic Features (*HF*), which have been used in some existing work, include Average Permalink/Sentence Length, Comment number, Organization Numbers, Opinion Rule, etc, which can be helpful for distinguishing some inclinations. In our approach, besides these heuristic features we also use the statistics of the presence of Cyber Words and Cyber Emoticons in feeds, which provides clues to personal and official feeds. Two public available cyberwords lists are used in our task, i.e., SmartDefine(*www.smartdefine.org*) including 465 words like "*IMNSHO* (which *means In My Not So Humble Opinion*)" and "*FAQ* (which means *Frequently Asked*

Questions)", and ChatOnline(*www.chatslang.com*), including 538 words like "*10x* (which means*Thanks*)" and "*Lemeno* (which means *Let me know*)". The integration these two acronyms lists is used as our acronyms lexicon to discover the acronyms in the tweets. Additionally, ChatOnline offers 273 commonly-used emoticons like "☺", "☹" and "^o^". These emoticons are used to detect the emoticon usage in blog posts, which is very practical in our experiments. Heuristic rules also defined to detect the word with repeating letters like *what'sssss up*", "*soooo guilty*" and "*Awesome!!!!*" etc. These repeating words usually show a strong emotion in the blog post. In case of *lexicon features*, we introduce the SentiWordNet(*sentiwordnet.isti.cnr.it*) and Wilson Lexicon (Wilson Theresa and Janyce Wiebe 2003), which are vital and commonly used in identifying opinionated inclination feeds. However, most of these features suit the opinionated inclination and may not work well in other inclinations, and may introduce noise to other inclinations, especially for the factual and shallow inclinations. Besides, because of employing two opinion lexicons, the feature structure is unbalanced for other facets. Thus, in order to overcome these defects and discover more faceted features, we take the effort to explore some useful features from corpora.

Here, we propose a feature expansion approach by **learning feature** candidates through feedback information of faceted feeds. The idea of learning feature is to introduce feature candidates from the first faceted ranking, and then learn some important feature candidates for new faceted ranking. Since TREC has released some annotated faceted feeds, it can be used as a criterion for feature learning. It is the advantage of this approach that it can learn useful features from feedback posts automatically. Different from inclination-dependent heuristic features, the learning process can be easily applied to any new inclination.

There are two steps to learn feedback features in feature expansion: feature candidate collections, and feature selection. In feature candidate collection, besides introducing the *unigram features* (*UF*), we consider the word associations, which we name as *pattern features* (*PF*). All high frequency unigrams are collected first as feedback feature candidates. However, for pattern features it is not feasible if we treat each pair of unigram as pattern candidates for the size of possible paired associations is nearly up to $4*10^8$ in our feature space. We need to measure the collaborative contribution of each pattern. Several information theoretic methods, such as Chi-square, log-likelihood ratio and mutual information, are applicable for this purpose. We choose Point-wise Mutual Information (PMI) for its easy implementation and relative performance, which are suitable for our massive dataset. The formula of PMI is as follows:

$$PMI(t_i, t_j) = log_2(P(t_i, t_j)/P(t_i)P(t_j))$$

where $P(t_i, t_j)$ is the joint probability that both unigrams appear in feeds, and $P(t_i)$ is the probability that a word *i* appears in feeds.

By now, we have collected unigram and pattern feature candidates, which are mainly opinion-independent ones and unbiased for particular inclinations, resulting in more balanced feature structure. A byproduct of feature expansion is that the unprocessed feature candidates contain too much useless information, which not only wastes computing resources but also harms the performance especially for massive dataset. For example, if all unigrams (more than 20,000) and pattern features (5000 selected with PMI) are selected as features in our experiments, it will take unpredictable time to extract the features from all feeds. Therefore, we need to select the top discriminative features with feature selection methods. There are several commonly used feature selection approaches. According to (Hua Liu and Hiroshi Motoda, 2007), information gain (IG) is

able to select both positive features and negative features. Thus, in our experiment IG is used to select features, and the formula is as follows:

$$IG(Ex, a) = H(Ex) - H(Ex|a)$$

where Ex is the set of all official annotated faceted feeds instances; $H(x)$ represents the entropy; $a \in Attr$, $Attr$ denotes all feature candidates including unigram and pattern features.

With lexicon-based features and feedback features, an unanswered question is how to determine the weight of both features. Though each opinion word has a polarity weight and a feature selection measure is assigned to each feedback features, these weights are not in the same scale. To unitize the weights of selected features, for each inclination we apply a SVM classifier with the default linear kernel and calculate the weight of a support vector from the trained model that corresponds to a feature. In linear kernel function, these weights stand for the coefficients of a linear function, and in certain degree they denote the importance of each support vector, which is the corresponding feature in our task. Eventually, the feeds are re-ranked with the sum of the products of the feature values and their weights.

3. Evaluation

The experiments are conducted on the blog08 collection(*ir.dcs.gla.ac.uk/test_collections*) crawled over 14 weeks. It contains permalinks, feeds, and related information. The size of the blog08 collection is up to 2.3TB. In order to efficiently handle the terabyte dataset and reduce noise, the raw dataset is first cleaned and filtered by several regular expression rules, e.g., removing unreadable text, filtering unnecessary HTML scripts, which reduce the size to 30% in total. Then, Indri is used to index the cleaned blog08 collection, and fetch the top 2000 related blog posts according to the 50 topics provided in TREC2009. Since the feeds are what the task is concerned with, we rank the feeds by summing the relevance scores of retrieved blog posts corresponding to the same feed number. The top 100 relevant feeds are obtained and evaluated in Table 1. TREC provides four measures: the mean average precision (MAP), R-Precision (rPrec), binary Preference (bPref) and Precision at 10 documents (P@10), among which MAP is the primary measure for evaluating the retrieval performance (TREC Blog Wiki, 2009).

Baseline Distillation	MAP	R-Prec	P@10	B-Pref
Language model	0.2494	0.3047	0.3590	0.2611
Official best	0.2756	0.2767	0.3206	0.3821
Official median	0.1752	0.1986	0.2447	0.3282
Official worst	0.0624	0.0742	0.0980	0.1410

TABLE 1 – Evaluation of baseline distillation compared with official best, median and worst

As shown in Table 1, our Indri-based language model ranks competitively against official submissions. Based on our baseline feed ranking, we conduct the faceted distillation. We first investigate 1500 opinion words from Lexicons of SentiWordNet and Wilson, about 20 K high-frequency (presence more than 5 times) unigram features are first collected from the top ranking five feeds as feature candidates. Then, we calculate the PMI of every pair of unigrams as a criterion of selecting more contributable pattern features. The top 5000 pattern features are heuristically selected as another source of feature candidates.

With the above feature candidates, IG is employed to select the features that contribute more. Instead of using all instances in the official released answers, we calculate $H(Ex|a)$ using the top five feeds in our experiments. This change can greatly reduce the complexity of computing and make our approach more adaptable for the massive data collection. The top five feeds are a good surrogate for the whole feed set as they are statistically found to contain an approximately equal number of faceted and non-faceted feeds. More importantly, this "shortcut approach" adapts very

well to the large dataset. We select the top ranked features for each inclination (examples are illustrated in Table 2). From the table, we also find that selected words really have the trend to express the meaning of the inclination, like "*argument*", "*rather*", "*powerful*" for opinionated. More important is that we observe these selected features, especially for pattern features, are topic related. For example, unigrams in Personal inclination, "*Fish*", "*boat*", "*river*" usually are related with topic 1189 "*personal travel stories*"; the pattern feature in Opinionated inclination "[synthesis, membrane]" usually present in ophthalmology treatment articles like topic 1194 "*macular degeneration*"; the pattern feature in In-depth inclination "[genealogy, ancestor]" is frequently related with topic 1155 "*2012 catastrophe armageddon predictions*". A similar observation is also found in (Yi Hu and Wenjie Li, 2010), which points out that topical terms and its context can be quite useful in determining the polarity. Thus this may indicates that the topic words and patterns are important for faceted re-ranking as well. Then, these selected features are also used to train the faceted models, and then the weights of these features can be inferred by the trained models. In practice, we use the same strategy to randomly divide the top five feeds into training and testing datasets (ratio 4:1). Then, the weights of support vectors are calculated from trained models as the weights of these features for facet re-ranking. With selected features and their weights, feeds are re-ranked according to each inclination, and for comparison, ranking without feedback features (HF+LF) is evaluated as well.

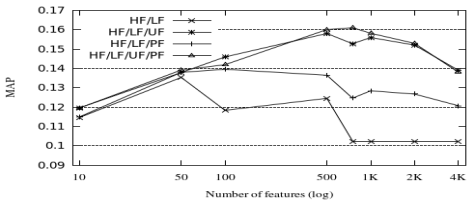

FIGURE 1 – Re-ranking with different number of features

Table 3 compares four rankings, re-ranking with heuristic and lexicon features [*HF+LF*]; re-ranking with heuristic, lexicon and unigram (top 500) features [*HF+LF+UF*]; re-ranking with heuristic, lexicon and pattern (top 500) features[*HF+LF+PF*] and re-ranking with heuristic, lexicon, unigram and pattern (top 750 features)[*HF+LF+UF+PF*]. The t-test is used to test whether a significant improvement can be achieved with feedback features. The T-values of [*HF+LF+UF*], [*HF+LF+PF*] and [*HF+LF+UF+PF*] are 4.78, 2.74 and 4.64, respectively, which are larger than $T_{0.05}$ = 1.76. This indicates that the re-rankings with feedback features achieve a significant improvement, and outperform the best of official runs. By using both unigram and pattern feedback features we obtain the best performance. From the evaluation of unigram features (*HF+LF+UF*) against without unigram features (*HF+LF*), we can find that great improvements are observed for factual, personal and official identification. It is thus plausible that those inclinations may be more amenable to the usage of words rather than some heuristic features. From the evaluation of pattern features (HF+LF+PF) against HF+LF, we can observe encouraging improvement on the shallow inclination, which may be the most difficult inclination for feature extraction, and this may also hint that the shallow inclination relies on word patterns more than single words, which coincides with our intuition that while a single word may be able to express opinionated or personal inclination, e.g., "rather", "better", "personally", it usually is hard to convey a shallow thought with only one word.

In the last experiment, comparisons are made to investigate the influence of different numbers of features selected. Figure 1 show four rankings mentioned above with different feature numbers. Obviously, the feedback feature is important for the faceted re-ranking. They peak at 750 features, 500 features and 100 features for re-ranking with the combination of unigram and pattern features, re-ranking with unigram features and re-ranking with pattern features, respectively. The flat tail of without-feedback approach (HF+LF) can be explained by the fact that only about 750 out of the 1500 features (shown by the points in the circle) are present in the features. We also notice that the pattern features have positive influence for faceted re-ranking, though there are only 0.83 percentage improvements in the point of 750. The bottom line shown in this figure is that re-ranking with feedback features outperforms that without feedback. This proves that feedback features are obviously effective in faceted blog distillation.

4. Conclusion

To sum up, feedback feature expansion coupled with feature selection is effective and efficient for faceted blog distillation and adapts well to the terabyte dataset. It helps to automatically discover relevant and discriminative features. Comparing with pattern features, unigram features play a more vital role in the tasks undertaken. In the future, we will investigate how to select more significant pattern features and use these pattern features to further improve the contribution.

Acknowledgements

The work presented in this paper is supported by a Hong Kong RGC project (No. PolyU 5230/08E).

	MAP						
	All	**Opinionated**	**Factual**	**Personal**	**Official**	**In-depth**	**Shallow**
Best 09	0.1261	0.1259	0.1350	0.1855	**0.1965**	0.1489	**0.1298**
HF+LF	0.1022	0.1340	0.0222	0.1754	0.1143	0.1859	0.0701
HF+LF+UF	0.1581	0.1467	0.1322	0.2166	**0.2333**	0.2091	0.0748
HF+LF+PF	0.1365	**0.1687**	0.1546	0.2067	0.1043	0.1294	0.0906
HF+LF+UF+PF	**0.1611**	0.1472	**0.1581**	**0.2351**	0.1860	**0.2210**	0.0918

TABLE 3 – Evaluation against each inclination

	Opinionated	**Factual**	**Personal**	**Official**	**In-depth**	**Shallow**
Unigram features	Political, national, Development, Maintain, Administration, Report, argument Powerful, chief	Election, target, agreement, status, Power, mission, Gravity, scientist, indicate	Fish, fly, catch, gear, Trout, boat, river, Lake, wait, sail	Breed, veterinarian, puppy, potty, groom, breed, purebred, bred	Ancestor, surgeon, surname, software, database, passenger, index, census	Learn, software, cosmetic, pocket, fine, surgeon, Procedure, religion, spiritual, tone
Pattern features	[Synthesis, membrane] [molecule, membrane] [metabolism, membrane]	[rocket, shuttle] [lunar, luna] [communist, missile] [thigh, underneath] [scalp, jaw]	[psychiatric, psychiatry] [marine, marina] [lunar, luna] [surgeon, surgery]	[veterinarian, veterinary] [veterinarian, rabbit] [dental, gum] [diarrhea, vomit]	[genealogy, ancestor] [genealogy, genealogist] [census, genealogy] [census, ancestor]	[genealogy, genealogist] [genealogy, ancestor] [psychiatric, psychiatry] [census, genealogy]

TABLE 2 – Examples of the selected Unigram and pattern features

References

Bouma, Gerlof. (2009). *Normalized* (*Pointwise*) *Mutual Information in Collocation Extraction*. In proceedings of the Biennial GSCL Conference, P31–40, Tübingen, Gunter Narr Verlag.

David Hannah, Craig Macdonald, Jie Peng, Ben He and Iadh Ounis. (2007). *University of Glasgow at TREC2007: Experiments in Blog and Enterprise Tracks with Terrier*. In proceedings of the 15th Text Retrieval Conference.

Hua Liu and Hiroshi Motoda. (2007). *Computational Methods of Feature Selection*. Chapman&Hall/CRC, P257-268, London.

Kiyomi Chujo, Masao Utiyama, Takahiro Nakamura and Kathryn Oghigian. (2010). *Evaluating Statitically-extracted Domain-specific Word Lists*. Corpus, ICT, and Language Education. Glasgow, UK.

Mejova Yelena and Thuc Viet Ha. (2009). *TREC Blog and TREC Chem: A View from the Corn Fields*, In proceedings of TREC09, University of Iowa.

Mostafa Keikha, Mark Carman, et al. (2009). *University of Lugano at TREC2009 Blog Track*. In proceedings of TREC09. Lugano, Swiss.

Richard McCreadie, Craig Macdonald and Iadh Ounis. (2009). *University of Glasgow at TREC2009: Experiments with Terrier*. In proceedings of TREC2009. Glasgow, Scotland, UK.

Seung-Hoon Na, Yeha Lee, Sang-Hyob Nam, and Jong-Hyeok Lee. (2009). *Improving Opinion Retrieval Based on Query-Specific Sentiment Lexicon*. ECIR, Berlin Heidelberg.

TREC Blog Wiki. 2009. Http://ir.dcs.gla.ac.uk/wiki-/TREC-BLOG.

Wilson, Theresa, and Janyce Wiebe. (2003). *Identifying opinionated sentences*. In proceedings of NAACL03, P33–34.

Wu Zhang and Clement Y. (2007). *UIC at TREC 2007 Blog Track*. In proceedings of the 15th Text REtrieval Conference.

Xuanjing Huang, Bruce Croft, et al. (2009). *Fudan University: A Unified Relevance Model for Opinion Retrieval*. ACM of Conference on Information and Knowledge Management, Hong Kong.

Yi Hu and Wenjie Li. (2010). *Document Sentiment Classification by Exploring description model of topical terms*. Computer Speech and Language 25(Jul. 2010), P386-403.