IJCLA VOL. 4, NO. 1, JAN-JUN 2013, PP. 175–194
RECEIVED 07/12/12 ACCEPTED 11/01/13 FINAL 12/03/13

Facet-Driven Blog Feed Retrieval

LIFENG JIA,[1] CLEMENT YU,[1] WEIYI MENG,[2] AND LEI ZHANG[1]

[1] *University of Illinois at Chicago, USA*
[2] *Binghamton University, USA*

ABSTRACT

The faceted blog distillation task retrieves blogs (i.e. RSS feeds) that are not only relevant to a query but also satisfy an interested facet. The facets under consideration are opinionated vs. factual, personal vs. official and in-depth vs. shallow. For the opinionated/factual facets, we propose a classifier that uses syntactic and semantic features to determine whether an opinion in blog documents is relevant to a given query. For the personal/official facets, we propose three classifiers that are learned based on different assumptions to categorize a blog document into either the personal or the official class. For the in-depth/shallow facets, we propose to calculate the depth of the coverage of a blog document on a given query by the occurrences of the concepts related to the query. Dependencies among different facets are also discussed. Experimental results on TREC Blogs06 and Blogs08 collections show that our techniques are not only effective in finding faceted blogs but also significantly outperform the best known results over both collections.

1 INTRODUCTION

Faceted blog distillation task is simply defined as: "*find me a quality blog with a principal, recurring interest in X*" [1]. Three pairs of quality aspects (called facets) of blogs are defined: 1)*Opinionated vs. Factual*: Some blogs convey opinionated comments on the topics of interest while others report factual information; 2)*Personal vs. Official*: Some blogs are written by individuals to depict their personal experiences while others

are written by companies to deliver their commercial influences; 3) *In-depth vs. Shallow*: Some blogs express in-depth thoughts and analysis on the reported issues while others simply provide quick bites on these topics without analyzing the implications of the provided information.

In this paper our aim is to achieve high retrieval effectiveness for faceted blogs, such as opinionated blogs. A blog (i.e. a RSS feed) consists of a set of blog documents (or called blog posts). We use the term *document* to represent a blog document (post) and the term *feed* to represent a blog. Faceted blog distillation can be seen as a two-phase task. Given a query Q and one of three pairs of facets mentioned above, 1) feeds (or documents) are ranked by only addressing their topical relevance to Q; 2) the feeds (or documents) from Phrase 1 as the baseline are re-ranked based on the pair of facets. Since TREC provided three baselines (i.e. the ranking of feeds in Phase 1), we only present the facet-finding techniques. There are three challenges discussed below for faceted blog distillation.

The retrieval of the opinionated blog documents is exactly the opinion retrieval problem [2]. It aims at retrieving the documents that convey the opinions relevant to a query. Since a blog document may contain opinions about multiple topics, the first challenge is how to capture opinions in a document that are related to the query. The state-of-the-art techniques are proximity-based [3, 4, 5, 6, 7]. If an opinion is close to the terms of a query within an blog document, it is likely to be relevant to the query. But the proximity-based determination is not sufficiently accurate, so we propose to use both the syntax and the semantics of a sentence to determine the opinion relevance. In addition, the query-referencing pronouns are identified by co-reference resolution and the key concepts (to be defined in Section 3) related to the query are extracted from knowledge bases. In this way, opinions not directly applicable to a query but applicable to those query-referencing pronouns or the key related concepts can be determined to be relevant to the query. Determining whether a blog document delivers the personal experiences or official information with commercial interests is the second challenge. To address this challenge, we study a research issue: should the personal or official facet of a blog document be independent of the query i.e. should a blog document be considered as a personal or official one irrespective of the query? This issue which has not been examined by other researchers has a direct impact on effectiveness. Moreover, we observe that people often express some opinions when describing their personal experiences. Thus we propose to examine whether the personal or official facet of a blog document is dependent on its opinionated or factual facet. The third challenge is

how to identify the in-depth or shallow blog documents. In-depth documents should provide in-depth thoughts and analysis about the queries. Since "thoughts" may be expressed as "opinions" and "analysis" means the depth of the coverage of blog documents, we explore these two characteristics in our solution.

This paper has the following contributions. (1) We propose a classifier to determine whether the opinions in blog documents are relevant to a given query. (2) We propose a set of classifiers to classify blog documents into personal or official classes. (3) We propose an approach to measure the in-depth or shallow facet of documents. Experiments show that the proposed techniques are effective.

2 Related Work

Besides the opinion retrieval studies, there is extensive research on opinion mining. Most opinion mining studies ignore the determination of the relevance of opinion and assume the opinions in their corpora (mainly product reviews) are always related to the object (product). They focus on how to relate an opinion to the different aspects of the object or to the opinion holder (who expresses the opinions). Instead of the opinion relevance to the opinion holder [8, 9, 10], our work studies the relevance of an opinion toward the object (mentioned in the query). The aspects in opinion mining roughly correspond to the key-related concepts in our work. The key differences between their works [11, 12, 13, 14, 15, 16, 17] and our work are: 1) the objects in their works are mainly products in reviews, while the objects mentioned in TREC queries come from diversified domains. Their techniques of mining the aspects of products are applied to product reviews and may not be applicable to the key related concepts of TREC queries over blog corpora. Therefore, we develop techniques to extract the key related concepts of query concepts from knowledge bases. 2) The utilization of key related concepts aims at recognizing the relevant opinions. Some relevant opinions in blog documents are not directed toward the objects (mentioned in the queries) but applicable to those key related concepts.

For finding the personal or official documents, some studies [18, 19, 20, 21] simply assume that the personal or official documents are the opinionated or factual documents respectively. Other studies calculate the personal or official facet scores based on dictionaries [22], heuristics [23, 24] and classifiers [25, 26]. No previous work studied our first research

issue: whether a document being personal or official is independent of the query.

To identify in-depth or shallow documents, the cross entropy between a blog document *d* and the whole collection is calculated as the in-depth score of *d* [20, 21]. Various heuristics [23, 22, 24] are proposed to measure the in-depth and shallow facets of documents. For example, an in-depth document is likely to be longer in terms of the number of terms than a shallow document. We propose to measure the depth of the coverage of a document *d* on a query topic by the occurrences of concepts in d which are closely related to the query.

3 OPINIONATED VS. FACTUAL

In this section, we introduce how to measure the extents of blog documents being opinionated and factual. Given a query Q and a blog document *d*, we first utilize an classifier [7] to classify the sentences in d into opinionated or factual ones. This classifier assigns each sentence an opinion or a factual score. Then, we determine whether the opinionated or factual sentences are relevant to Q. Finally, we calculate the opinionated (or factual) facet score of d is the sum of the opinion (or factual) scores of the relevant opinionated (or factual) sentences.

The key is how to recognize the opinionated/factual sentences relevant to Q. For each opinionated sentence s, we decide s is relevant to Q if the following two conditions are satisfied. The first condition is that s and Q co-occur within a window of five sentences consisting of s, two proceeding ones and two succeeding ones [7]. But this proximity-based condition alone is not sufficient to accurately determine the relevance of s to Q. Therefore, we stipulate a second condition to further determine whether s is indeed relevant to Q. Specifically, we first identify the occurrences of Q in s, then resolve the query-referencing pronouns in s and finally identify the hypernyms of Q or the key related concepts of Q in s, if present. We denote the occurrences of Q, the query-referencing pronouns and the hypernyms and the key related concepts of Q as *target terms*. We also identify the opinion terms in s by two opinion lexicons [27, 28]. The second condition is whether s has an opinion term related to one of the target terms. If *s* contains no target terms, the opinion in *s* is irrelevant to Q, in spite of its close proximity to Q. The hypernyms of Q and the key-related concepts of Q are essential as illustrated. For example, given Q = "*Brokeback Mountain*", the opinion terms that are related to a hypernym

Table 1. A Sample of *Syntactic* (italics) and **Semantic** (bold) Features.

Feature Name	Feature Description (*O* = Opinion Term, *T* = a Target Term)
TSub	Valued *TRUE* when *T* is the subject (*Sub*) of the opinionated sentence;
OPred	Valued *TRUE* when *O* is the predicate (*Pred*) of the opinionated sentence;
OModNNT	Valued *TRUE* when *O* modifies a noun *N*; *T* and *N* satisfy the following condition: *T* is a non-person concept but *N* is the hyponym of person or vice versa;
SpecialPhrase	Valued *TRUE* when *O* forms some special phrases without opinions, such as "as **well** as";

of Q, "*movie*" or a key related concept of Q, "*Health Ledger*" can represent the relevant opinions to Q. Factual sentences do not have "factual terms" to signify their factualness as there is no "factual lexicon". So we determine a factual sentence s to be relevant to Q, if s and Q co-occur within a window of five sentences.

Query-Referencing Pronouns, Hypernyms and Key Related Concepts. To determine whether an opinionated sentence s is relevant to a query Q, at least one opinion term in s is related with Q. Some opinion terms that are not directly related with Q but related with the pronouns referencing Q can convey the opinions relevant to Q. Specifically, Illinois Coreference toolkit [29] is used on the paragraph containing s to resolve the pronouns referencing Q. Besides the pronouns, the opinion terms related with the hypernyms or the key related concepts of queries are relevant to the queries. Specifically, key concepts are related to Q by the "part-of" and "equivalence" relationships. There are other possible relationships, such as the "associative" relationship between two concepts. However, in our opinion, they are unsuitable for determining the opinion relevance toward the query. We use three knowledge bases: YAGO [30], DBPedia[3] and Freebase[4], to extract the hypernyms and the key related concepts of queries. The hypernyms of a query Q can be automatically identified by their associations with Q by the relationships 'IsA" in YAGO, "type" in DBPedia or "category" in Freebase. But the key related concepts cannot be directly extracted from the knowledge bases, because relationships in these knowledge bases are defined in free-text and determining which free-text relationships correspond to the "part-of" and the "equivalence" relationships is difficult. We manually examine the free-text relationships in these knowledge bases to determine whether they can simulate either the "part-of" or the "equivalence" relationship. For example, given a relationship, "starring", two concepts, "*Leonardo DiCaprio*" and "*Titanic*" are associated by "starring" in the knowledge bases. "*Leonardo*

[3] http://wiki.dbpedia.org/Ontology

[4] http://www.freebase.com/

DiCaprio" can be considered as a part of movie "*Titanic*", so "starring" is determined to be qualified for simulating the "part-of" relationship. Following the selection criteria above, a list of 313 relationships (10 from YAGO, 104 from DBPedia and 199 from Freebase) is established. Note that the manual examination of relationships is carried out only once before the query processing. No query is involved in the examination. Given such a list of relationships, the key-related concepts of any query can be retrieved from these knowledge bases automatically.

Syntactic and Semantic Features. Given an opinionated sentence s, after all the target terms are identified, if present, we determine whether an opinion term O is related to one of the target terms T in terms of s's syntax and semantics. We treat the relevance of O to T within s as a classification problem. We propose a set of features based on syntax and semantics. Table 1 presents a sample of the proposed features and those features described below are excluded. The syntax of a sentence can be expressed by typed dependencies and a parse tree, both of which are obtained by Stanford parser [31]. We propose typed dependency (TD) features and (parse) tree node (TN) features. *Typed Dependency*: given the TDs of an opinionated sentence, an undirected TD graph is built where the vertices are the terms and the edges are TDs between terms. A TD path between term *A* and term *B* is a sequence of TDs between vertex *A* and vertex *B*. Given the shortest TD path SP between the opinion term and a target term, for each TD t_d in SP, we prefix t_d's name with SP's length and suffix t_d's name with its sequential position in SP. It is a TD feature. *Tree Node*: given a parse tree of an opinionated sentence, we ignore the directions of tree edges. We then find the shortest path SP from a leaf node A representing an opinionated term to a leaf node B representing a target term. We represent SP by a sequence of intermediate tree nodes by excluding A and B. For each tree node t_n in SP, we prefix t_n's name with SP's length and suffix t_n's name with its sequential position in SP. It is a TN feature. A short distance between an opinion term and a target term in a TD graph or in a parse tree indicates relevance of opinion.

Moreover, the Boolean features in Table 1 can indicate the relevance of the opinion terms to the target terms within an opinionated sentence. For example, a syntactic feature named *OTDiffC* is valued true when an opinion term and the target terms occur in different clauses, which indicates that they are unlikely to be related. We propose some semantic features too. For example, a semantic feature named *Comparison* is valued true when the opinionated sentence is a comparative or superlative one. The intuition is that an opinion in such a sentence is always directed

toward all entities involving the comparison and thus the opinion term is likely to be related to the target terms.

We sample 1108 training examples from TREC Blogs06 collection w.r.t. 50 TREC 2006 queries. Each example is a triple consisting of an opinion term, a query and an opinionated sentence containing them. The opinion term is manually labeled to be either relevant or irrelevant to the query. The query-referencing pronouns, hypernyms and key related concepts if present are identified. An opinion relevance classifier is trained by using the training data and the features.

4 PERSONAL VS. OFFICIAL

In this section, we present three classifiers. Each of them classifies the blog documents into either the personal or the official class. These classifiers examines the following two research issues. First, is the class of a document (personal vs. official) independent of the query i.e. should a document be considered as personal or official irrespective of the query? Second, is the class of a document dependent on whether the document is opinionated or factual? To build classifiers, a set of features and the training data are essential. TREC relevance judgements are used as the training data but they only provide facet judgments on feeds, instead of documents. Table 2 shows a sample of proposed features. The proposed features can be generally categorized into query independent ones (QID and QIF groups) and query dependent ones (QDD and QDF groups). The answers to these two issues influence the feature selection and the usage of the training data. Our proposed features can be partitioned into two classes: query-independent and query-dependent. Each class can be further partitioned into two subclasses: document level or feed level. These 4 subclasses are sketched below.

1) *Query Independent Document Level Features* (QID). A document can show some clues of its personal or official facet. For example, people are more interested in commenting the personal documents than the official ones. Thus the number of comments in a document is a good indicator of its personal or official facet. The more comments a document has, the more likely it is personal. In TREC Blogs08 collection, the average number of comments per document in personal feeds is 4.9 while that of official feeds is 1.1. We propose 22 QID features.

2) *Query Dependent Document Level Features* (QDD). An example feature is the number of sentences that are classified to be opinionated relevant ones to a given query. We propose 8 QDD features.

Table 2. A Sample of Features for Personal or Official Classification.

Group	ID	Feature Description (d = document, f = the feed containing d)	#
QID	D_1	No. of images in d;	1
QID	D_2	No. of sentences in d classified to be opinionated and the sum of their opinion scores;	2
QDD	D_3	No. of query terms in the title of d;	1
QDD	D_4	Similar to D_2, except the classified opinionated and relevant sentences to a given query are utilized;	2
QIF	F_1	The mean and the standard deviation of the feature D_1 of documents in f;	2
QDF	F_2	The mean and the standard deviation of the feature D_3 of documents in f;	2

3) *Query Independent Feed Level Features* (QIF). An example feature is the percentage of documents in a feed that have no first person pronouns. A higher percentage more likely indicates an official feed. We propose 51 QIF features.

4) *Query Dependent Feed Level Features* (QDF). An example feature is the percentage of documents in a feed whose titles contain at least one query term. We propose 3 QDF features

Three personal/official (PS/OF) classifiers are built based on different assumptions about those two research issues. Accordingly, three PS/OF modules are constructed. Each module uses a classifier and ranks the documents as below.

1) *Query Independent with Opinionated and Factual Features* (QIOPFT): By assuming that a document being personal or official is independent of queries but depends on its opinionated or factual facet, the first classifier QIOPFT is built as follows. Given a labeled feed *f*, all documents in *f* are used as the training data and they are assigned the same facet label as that of *f*. All query-independent features (QID and QIF groups) are utilized. After QIOPFT is learned over the training data by those features, a module using QIOPFT is established. In this module, a document *d* is first classified into either the personal or the official class. Then *d* is assigned by QIOPFT a classification score $PS(d)$ (or $OF(d)$), if it is classified into the personal (or official) class. Let $F_{OP}(d)$ and $F_{FT}(d)$ be the opinionated and factual facet scores of *d* respectively. Since we assume that the class of a document depends on its being opinionated or factual, the module using QIOPFT assigns the personal facet score, $F_{PS}(d)$, and the official facet score, $F_{OF}(d)$ of the document *d* as follows. Here, λ is empirically learned:

$$\begin{aligned} F_{PS}(d) &= \lambda F_{OP}(d) + (1-\lambda)PS(d), \\ F_{OF}(d) &= \lambda F_{FT}(d) + (1-\lambda)OF(d). \end{aligned} \tag{1}$$

2) *Query Dependent with Opinionated and Factual Features* (QDOPFT). By assuming that a document being personal or official is not only dependent on queries but also dependent on its opinionated or factual facet, the second classifier QDOPFT is built as follow. Given a labeled feed f, only the subset of documents that contains at least a query concept is considered to inherit the label of f. These query-dependent documents are utilized as the training data. QDOPFT is trained over this subset of training data but involves all features including both query-independent and query-dependent features (QID, QIF, QDD and QDF groups). Due to the assumption that the class of a document d depends on its being opinionated or factual, the module using QDOPFT assigns d a facet score by Equation (1) too.

3) *Query Independent without Opinionated and Factual Features* (QIwoOPFT). To train the third classifier QIwoOPFT, we make the assumption that a document being personal or official is independent of not only the query but also its opinionated or factual facet. QIwoOPFT is constructed using the same training data as the first classifier. However, the features used by OIwoOPFT are those query independent features (QID and QIF groups) with the exclusion of those features which are calculated based on the opinionated or factual sentences of documents, such as D_2 in Table 2. After QIwoOPFT is constructed, a document d is first categorized into the personal or the official class and is then assigned a classification score by QIwoOPFT accordingly. Due to the independent assumption between the personal/official facets and the opinionated/factual facets, we just use the classification score of d as its corresponding facet score.

We build QIOPFT and QDOPFT by the different assumptions about whether the class (personal or official) of documents is independent of queries, so we can answer the first issue by comparing their effectiveness. Experimental results in Section 7 show that QIOPFT yields better effectiveness than QDOPFT and we conclude that the class of documents is independent of queries. Acknowledging such a conclusion, we build QIOPFT and QIwoOPFT by the different assumptions about whether the class of documents depends on its opinionated or factual facet. We can answer the second research issue by comparing their effectiveness.

5 IN-DEPTH VS. SHALLOW

In this section, we present our techniques for in-depth and shallow facets. Intuitively, an in-depth analysis about a query Q should not only contain

Q, but also contain the related concepts of Q. So we propose an approach that identifies the related concepts of Q. The method is described in two steps: 1) Given the Wikipedia entry of each concept of Q, collect the anchor texts and the noun phrases in the subtitles as the candidates of the related concepts. 2) Calculate the association between a candidate *e* and Q by Pointwise Mutual Information [32]; $P(e, Q)$ is the co-occurrence probability of e and Q. $P(e)$ (or $P(Q)$) is the occurrence probability of e (or Q). They are estimated by Google.

$$PMI(e, Q) = \log\left(\frac{P(e, Q)}{P(e)P(Q)}\right) \tag{2}$$

We propose two methods to measure the in-depth (or shallow) facet score of a document *d*, $F_{ID}(d)$(or $F_{SW}(d)$). The first method computes $F_{ID}(d)$ or $F_{SW}(d)$ without considering whether d is opinionated or factual. It assumes that d is a in-depth document if it provides deep analysis; otherwise, d is a shallow document. Let $RC(Q)$ be the top k (k = 30 in this paper) related concepts of Q and $CNT(e, d)$ be the count of *e* in *d*. $F_{ID}(d)$ and $F_{SW}(d)$ are calculated as below.

$$\begin{aligned} F_{ID} = Dep(d) &= \sum_{e \in RC(Q)} CNT(e, d) \cdot PMI(e, Q), \\ F_{SW}(d) &= 1 - Dep(d). \end{aligned} \tag{3}$$

The second method assumes that an in-depth document is likely to be opinionated and provides deep analysis; a shallow document is likely to be factual and provides no deep analysis. Let $F_{OP}(d)$ and $F_{FT}(d)$ be the opinionated and factual facet scores of d. After $Dep(d)$ score is normalized between 0 and 1, $F_{ID}(d)$ and $F_{SW}(d)$ can be alternatively calculated as below. λ is empirically learned.

$$\begin{aligned} F_{ID}(d) &= \lambda F_{OP}(d) + (1 - \lambda) Dep(d), \\ F_{SW}(d) &= \lambda F_{FT}(d) + (1 - \lambda)(1 - Dep(d)). \end{aligned} \tag{4}$$

6 AGGREGATION MODULE

In this section, we propose an aggregation method to calculate the facet score of each feed by the facet scores of its documents. Let Q be a query topic and D_Q be the set of documents retrieved by a topical retrieval system w.r.t. Q; given a feed *f*, D_f is the set of the documents in *f*; $IR(d)$

and $F_t(d)$ are the ad-hoc score of a document *d* from the topical retrieval system and the facet score of *d* for the facet *t* respectively. t is one of six facets discussed. In this paper we use TREC baselines to obtain $IR(d)$ and D_Q. For any feed *f*, its ad-hoc score, $IR(f)$ and its facet score, $F_t(f)$, are calculated as below:

$$IR(f) = \frac{|D_Q \cap D_f|}{|D_f|} \cdot \sum_{d \in D_Q \cap D_f} IR(d), \tag{5}$$

$$F_t(f) = \frac{|D_Q \cap D_f|}{|D_f|} \cdot \sum_{d \in D_Q \cap D_f} F_t(d)$$

An aggregated score $AS_t(f)$ is computed as below. All feeds are ranked in descending order of their aggregated scores. Here, α is empirically learned:

$$AS_t(f) = \alpha IR(f) + (1 - \alpha) F_t(f) \tag{6}$$

7 EXPERIMENTS

Experimental Setup. Since opinion retrieval plays a central role in faceted blog distillation, we first evaluate our opinion-finding techniques using 100 TREC 2007-2008 queries over five TREC baselines (of documents) from TREC Blogs06 collection. The performance metrics are Mean Average Precision (MAP), R-Precision (R-Prec), binary Preference (bPref) and Precision at top 10 documents (P@10). MAP is the most important metric. Another set of experiments is designed to evaluate the proposed facet-finding techniques using 70 TREC 2009-2010 queries over three TREC baselines (of feeds) from TREC Blogs08 collection. These 70 queries consist of 20 queries with opinionated/factual facets, 18 queries with personal/official facets and 32 queries with in-depth/shallow facets. TREC Blogs08 collection is the only official blog collection for faceted blog distillation. MAP as the most important metric in TREC 2009-2010 is used here.

Opinion Retrieval Evaluation. Our opinion-finding technique is characterized by three sub-techniques: 1) the syntax and semantics features, 2) the hypernyms and the key related concepts from knowledge bases and 3) co-reference resolution for identifying query-referencing pronouns. We evaluate their impacts individually as follows.

We first use the opinion retrieval system [7] as baseline. It determines the opinion relevance to a query only based on the proximity condition

of five sentences. Denote it by System I. Then, we configure a second system (denoted by System II) that in addition to the proximity condition, employs the proposed classifier to further determine the relevance of opinionated sentences. It uses the syntactic and semantic features but the hypernyms and the key related concepts of the query concepts and the query-referencing pronouns are not identified. The target terms are only the query concepts. The third system (denoted by System III) uses not only the classifier but also the hypernyms and the key related concepts as target terms. Co-reference resolution is not used. The fourth system employs all three sub-techniques and performs co-reference resolution by Illinois Co-reference toolkit [29]. Denote it by System IV.

All systems are given the same ad-hoc baseline obtained by the topical retrieval system [33] as input and re-rank the documents by addressing the opinionated facet. Since 50 TREC 2006 queries are used for training the opinion relevance classifier, all systems are evaluated by 100 TREC 2007-2008 queries. Table 3 shows their performance. System II achieves statistically significant improvements over System I in all measures. It indicates the classifier that is based on syntax and semantics is effective in determining the opinion relevance, even though only query concepts are used as target terms. The utilization of the hypernyms and the key related concepts in System III contributes to consistent improvements over System II in all measures. Specially, the improvements in MAP and bPref are statistically significant. These improvements indicate that the utility of the hypernyms and the key related concepts are beneficial for determining the opinion relevance. In comparing System IV with System III, the resolution of pronouns contributes to the marginal improvements in all measures. We employed a different co-reference resolution toolkit OpenCalais[5] without observing significant performance difference.

Overall, System IV achieves statistically significant improvements over System I in all measures. It indicates the proposed techniques together are very effective. In addition, we compare System VI with the state-of-the-art opinion retrieval method called laplaceInt [3]. It determines the relevance of opinions to queries by their proximities, achieving the best performance over five TREC baselines from TREC Blogs06 collection by using 50 TREC 2008 queries. We evaluate System IV over those five baselines by the same set of queries and compare its performance with that of laplaceInt. Table 4 shows that System IV consistently and significantly outperforms laplaceInt over these five baselines in MAP,

[5] http://www.opencalais.com

Table 3. Comparison of System K with System K-1 (K = 2, 3, 4); $\triangle$ denotes statistically significant improvements over System K-1 by System K at $p < 0.05$; ▲ denotes statistically significant improvements over System I by System IV at $p < 0.05$.

	MAP	R-Prec	bPref	P@10
System I	0.4304	0.4497	0.4790	0.6560
System II	$0.4771^{\triangle}$	$0.4875^{\triangle}$	$0.5091^{\triangle}$	$0.7060^{\triangle}$
System III	$0.4835^{\triangle}$	0.4900	$0.5188^{\triangle}$	0.7100
System IV	$\mathbf{0.4843}^{\blacktriangle}$	$\mathbf{0.4904}^{\blacktriangle}$	$\mathbf{0.5192}^{\blacktriangle}$	$\mathbf{0.7120}^{\blacktriangle}$

R-Prec and bPref. For P@10, System IV outperforms laplaceInt by 5.0% averagely.

Faceted Blog Distillation Evaluation. We now evaluate all proposed faceted-finding techniques over the three TREC baselines from TREC Blogs08 collection. In addition, we compare the performance of our techniques with the best performance in TREC 2010 [34]. They are the "hitFeeds" runs [26] and the "LexMIRuns" runs [20]. Note that the parameters λ and α (from Equations (1), (4) and (6)) are learned as follows. We try all possible values for λ and α from 0.1 to 1.0 with interval of 0.1 respectively. The values of λ and α that perform best for TREC 2009 queries are used to evaluate TREC 2010 queries and vice versa. Moreover, the opinion relevance classifier used in this set of experiments is trained by using TREC 2006 queries while tested by 70 TREC 2009-2010 queries.

Opinionated and Factual Effectiveness. Tables 5 and 6 show the evaluation of our opinionated (OP) and factual (FT) blog distillation method (denoted by OPFT) over three baselines by using 20 TREC queries with OP and FT facets. OPFT consistently achieves significant improvements in both facet performance over all three baselines. We also compare OPFT with the state-of-the-art methods.

Tables 5 and 6 show that OPFT consistently and significantly outperforms the best performance in both facet performance. Xu et al. [35] only studied opinionated blog distillation by using those 20 TREC queries over the same baselines. Our performance outperforms theirs by 4.0% in mean MAP score. We show the average improvement without showing their results due to space limit.

Personal and Official Effectiveness. We evaluate three proposed personal (PS) and official (OF) blog distillation methods by 18 TREC queries with PS and OF facets over the three baselines. The three methods use

Table 4. Comparison of System IV with laplaceInt; ▲ denotes statistically significant improvements over baselines by System IV at $p < 0.05$.

	MAP	R-Prec	bPref	P@10
baseline1	0.3239	0.3682	0.3514	0.5800
laplaceInt	0.4020	0.4412	0.4326	0.6920
System IV	**0.4294**▲	**0.4610**▲	**0.4631**▲	**0.7260**▲
baseline2	0.2639	0.3145	0.2902	0.5500
laplaceInt	0.2886	0.3411	0.3166	0.5860
System IV	**0.3526**▲	**0.4108**▲	**0.3862**▲	**0.6400**▲
baseline3	0.3564	0.3887	0.3677	0.5540
laplaceInt	0.4043	0.4389	0.4247	**0.6660**
System IV	**0.4192**▲	**0.4447**▲	**0.4374**▲	**0.6660**▲
baseline4	0.3822	0.4284	0.4112	0.6160
laplaceInt	0.4292	0.4578	0.4485	**0.7140**
System IV	**0.4540**▲	**0.4836**▲	**0.4811**▲	0.7040▲
baseline5	0.2988	0.3524	0.3395	0.5300
laplaceInt	0.3223	0.3785	0.3715	0.6120
System IV	**0.3535**▲	**0.4015**▲	**0.3944**▲	**0.6860**▲

three proposed PS/OF classifiers and are named as QDOPFT, QIwoOPFT and QIOPFT respectively. Note that TREC 2009 query topics are tested over the PS/OF classifiers that are trained over the relevance judgments of TREC 2010 query topics and vice versa. The comparison between QDOPFT and QIOPFT can answer our first research issue: "Is a document being personal or official independent of the query?" Tables 5 and 6 show that QIOPFT outperforms QDOPFT in terms of the mean MAP score of PS and OF performance over the three baselines. So we believe that a document being personal or official is independent of the query. To address the second research issue: "Is a document being personal or official dependent on whether it is opinionated or factual?", we conduct the comparison between QIwoOPFT and QIOPFT.

Tables 5 and 6 show that the QIOPFT consistently outperforms the QIwoOPFT over three baselines in terms of PS and OF facet performance. So we conclude that a document being personal or official is dependent on its opinionated or factual nature. Since a document being personal or official is independent of the query, a possible concern is that there may be feeds that are judged to be personal (or official) for some TREC 2009 queries and they have the same judgments for some TREC 2010 queries. This may cause our classifiers to be overfitting, because the

Table 5. Performance of faceted blog distillation modules, part 1. ▲ (▼) and △ (▽) denote statistically significant improvements (deteriorations) at $p < 0.05$ and $p < 0.1$.

	Opinionated Facet Effectiveness		
	BASELINE 1	BASELINE 2	BASELINE 3
baseline	0.2426	0.1318	0.1001
OPFT	**0.2742**$^{(13.0\%)}$	**0.1721**$^{(30.6\%)}$	**0.1533**$^{(53.1\blacktriangle)}$
hitFeeds	0.2436	0.1319	0.1015
LexMIRuns	0.2518	0.1428	0.1199
OPFT	**0.2742**$^{(12.6\%,8.9\%\triangle)}$	**0.1721**$^{(30.5\%\triangle,20.5\%)}$	**0.1533**$^{(51.0\%\blacktriangle,27.9\%)}$

	Personal Facet Effectiveness		
	BASELINE 1	BASELINE 2	BASELINE 3
baseline	0.2097	0.1527	0.0895
QDOPFT	0.2444$^{(16.5\%)}$	0.1667$^{(9.2\%)}$	0.1677$^{(87.3\%\triangle)}$
QIwoOPFT	0.1751$^{(-16.5\%,N/A)}$	0.1167$^{(-23.6\%,N/A)}$	0.0872$^{(-2.6\%)}$
QIOPFT	0.2440$^{(16.4\%,-0.2\%,39.3\%)}$	**0.1966**$^{(28.7\%,17.2\%\triangle,68.5\%\blacktriangle)}$	**0.1683**$^{(88.0\%\triangle,0.4\%,93.0\%\blacktriangle)}$
hitFeeds	0.2126	0.1533	0.0911
LexMIRuns	**0.2727**	0.1607	0.0875
QIOPFT	0.2440$^{(14.8\%,-10.5\%)}$	**0.1966**$^{(28.2\%,22.3\%)}$	**0.1683**$^{(84.7\%\triangle,92.3\%\blacktriangle)}$

	In-depth Facet Effectiveness		
	BASELINE 1	BASELINE 2	BASELINE 3
baseline	0.3298	0.2185	0.1616
IDSW	0.3283$^{(-0.5\%)}$	0.2357$^{(7.8\%)}$	0.2097$^{(29.8\%\triangle)}$
IDSWOPFT	**0.3339**$^{(1.2\%,1.7\%)}$	**0.2421**$^{(10.8\%,2.7\%\triangle)}$	**0.2141**$^{(32.5\%\blacktriangle,2.1\%)}$
hitFeeds	0.3300	0.2184	0.1643
LexMIRuns	0.3311	0.2185	0.1616
IDSWOPFT	**0.3339**$^{(1.2\%,0.8\%)}$	**0.2421**$^{(10.9\%,10.8\%)}$	**0.2141**$^{(30.3\%\blacktriangle,32.5\%\blacktriangle)}$

PS/OF classifiers are trained over the facet-judged feeds for TREC 2010 queries and then tested by TREC 2009 queries and vice versa. However, after examining the facet-judged feeds of 18 TREC PS/OF queries, the set of 181 facet-judged feeds for 8 TREC 2009 PS/OF queries and the set of 205 facet-judged feeds for 10 TREC 2010 PS/OF queries are disjoint.

Table 6. Performance of faceted blog distillation modules, part 2. ▲ (▼) and △ (▽) denote statistically significant improvements (deteriorations) at $p < 0.05$ and $p < 0.1$.

	Factual Facet Effectiveness		
	BASELINE 1	BASELINE 2	BASELINE 3
baseline	0.2520	0.1911	0.1419
OPFT	**0.2802**$^{(11.2\%)}$	**0.2032**$^{(6.3\%)}$	**0.1639**$^{(15.5\%\blacktriangle)}$
hitFeeds	0.2529	0.1913	0.1417
LexMIRuns	0.2477	0.1942	0.1622
OPFT	**0.2802**$^{(10.8\%,13.1\%)}$	**0.2032**$^{(6.2\%,4.6\%)}$	**0.1639**$^{(15.6\%\blacktriangle,1.0\%)}$

	Official Facet Effectiveness		
	BASELINE 1	BASELINE 2	BASELINE 3
baseline	0.2673	0.1957	0.2016
QDOPFT	0.2570$^{(-3.9\%)}$	0.2046$^{(4.5\%)}$	0.2102$^{(4.3\%)}$
QIwoOPFT	0.2658$^{(-0.6\%,N/A)}$	0.1674$^{(-14.5\%,N/A)}$	0.2085$^{(3.4\%,N/A)}$
QIOPFT	0.2690$^{(0.6\%,4.7\%,0.6\%)}$	**0.2449**$^{(25.1\%,19.7\%,46.3\%\blacktriangle)}$	**0.2366**$^{(17.4\%,12.3\%,13.5\%)}$
hitFeeds	**0.2700**	0.1957	0.1985
LexMIRuns	0.2662	0.1882	0.2016
QIOPFT	0.2690$^{(-0.4\%,1.0\%)}$	0.2449$^{(25.1\%,30.1\%)}$	**0.2366**$^{(19.2\%,17.4\%)}$

	Shallow Facet Effectiveness		
	BASELINE 1	BASELINE 2	BASELINE 3
baseline	0.1370	0.1125	0.0921
IDSW	**0.1450**$^{(5.8\%\triangle)}$	**0.1293**$^{(14.9\%)}$	0.1043$^{(13.2\%)}$
IDSWOPFT	0.1421$^{(3.1\%,-0.2\%)}$	0.1289$^{(14.6\%,-0.3\%)}$	**0.1144**$^{(24.2\%,9.7\%)}$
hitFeeds	0.1378	0.1123	0.0913
LexMIRuns	0.1279	0.1046	0.0910
IDSWOPFT	0.1421$^{(3.1\%,11.1\%)}$	0.1289$^{(14.8\%,23.2\%)}$	**0.1144**$^{(25.3\%,25.7\%)}$

QIOPFT is the most effective and robust one among all three methods. So we compare its performance with the best known performance. QIOPFT significantly outperforms the "hitFeeds" runs and the "LexMIRuns" runs in both faceted performance. Gerani et al. [18] only studied personal blog distillation. We perform experiments using their queries and outperform their results by 18.1% in MAP. We show the average improvement without showing their results due to space limit.

In-depth and Shallow Effectiveness. We evaluate our in-depth (ID) or shallow (SW) methods by using 32 TREC queries with ID and SW facets. We first configure a method where the facet scores are calculated by Equation (2). The depth of documents are measured by the extent of the occurrences of related concepts to queries. Let IDSW denote this method. We then configure another method where the facet scores are calculated by Equation (4). It considers the depth or shallowness of a document not only by the related concepts but also by the OP or FT facet scores. Let IDSWOPFT denote this method. Tables 5 and 6 show that IDSW significantly improve the baselines in the ID and SW performance, which indicates the effectiveness of the usage of related concepts of queries to measure the depth of blog documents. IDSWOPFT is more robust and more effective than IDSW, because it not only outperforms IDSW in terms of the mean MAP scores for ID and SW performance but also consistently and significantly improves all three baselines in ID and SW performance. Thus, we believe that an in-depth document is likely to contain opinionated contents and a shallow document is likely to be factual. We observe that IDSWOPFT consistently and significantly outperforms those best performance in both faceted performance.

8 CONCLUSION

In this paper, we proposed techniques to classify and rank facet-oriented feeds. Moreover, we carefully studied a number of research issues in the construction of the classifiers. Some of these issues have not been addressed by earlier researchers. We set up different experiments to answer these research issues. Experiments demonstrated that our facet-finding techniques not only consistently outperform the three TREC baselines but also outperform the best results.

REFERENCES

[1] Macdonald, C., Ounis, I., Soboroff, I.: Overview of the trec 2009 blog track. In: TREC. (2009)
[2] Ounis, I., de Rijke, M., Macdonald, C., Mishne, G., Soboroff, I.: Overview of the trec-2006 blog track. In: TREC'06. (2006)
[3] Gerani, S., Carman, M.J., Crestani, F.: Proximity-based opinion retrieval. In: SIGIR '10. (2010)

[4] Santos, R.L.T., He, B., Macdonald, C., Ounis, I.: Integrating proximity to subjective sentences for blog opinion retrieval. In: ECIR '09. (2009)
[5] Vechtomova, O.: Facet-based opinion retrieval from blogs. Inf. Process. Manage. **46**(1) (January 2010) 71–88
[6] Zhang, M., Ye, X.: A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In: SIGIR '08. (2008)
[7] Zhang, W., Yu, C., Meng, W.: Opinion retrieval from blogs. In: CIKM '07. (2007)
[8] Breck, E., Choi, Y., Cardie, C.: Identifying expressions of opinion in context. In: IJCAI'07. (2007)
[9] Choi, Y., Breck, E., Cardie, C.: Joint extraction of entities and relations for opinion recognition. In: EMNLP '06. (2006)
[10] Johansson, R., Moschitti, A.: Reranking models in fine-grained opinion analysis. In: COLING'10. (2010)
[11] Ding, X., Liu, B.: Resolving object and attribute coreference in opinion mining. In: COLING '10. (2010)
[12] Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In: WSDM '08. (2008)
[13] Hu, M., Liu, B.: Mining and summarizing customer reviews. In: KDD '04. (2004)
[14] Joshi, M., Penstein-Rosé, C.: Generalizing dependency features for opinion mining. In: ACL-IJCNLP '09. (2009)
[15] Kobayashi, N., Inui, K., Matsumoto, Y.: Extracting aspect-evaluation and aspect-of relations in opinion mining. In: EMNLP-CoNLL'07. (2007)
[16] Popescu, A.M., Etzioni, O.: Extracting product features and opinions from reviews. In: HLT-EMNLP '05. (2005)
[17] Wu, Y., Zhang, Q., Huang, X., Wu, L.: Phrase dependency parsing for opinion mining. In: EMNLP '09. (2009)
[18] Gerani, S., Keikha, M., Carman, M., Crestani, F.: Personal blog retrieval using opinion features. In: ECIR'11. (2011)
[19] Jia, L., Yu, C.T.: Uic at trec 2010 faceted blog distillation. In: TREC. (2010)
[20] Keikha, M., Mahdabi, P., Gerani, S., Inches, G., Parapar, J., Carman, M.J., Crestani, F.: University of lugano at trec 2010. In: TREC. (2010)
[21] Zhou, Z., Zhang, X., Vines, P.: Rmit at trec 2010 blog track: Faceted blog distillation task. In: TREC. (2010)

[22] Li, S., Li, Y., Zhang, J., Guan, J., Sun, X., Xu, W., Chen, G., Guo, J.: Pris at trec 2010 blog track: Faceted blog distillaton. In: TREC. (2010)
[23] Guo, L., Zhai, F., Shao, Y., Wan, X.: Pkutm at trec 2010 blog track. In: TREC. (2010)
[24] Mejova, Y., Ha-Thuc, V., Foster, S., Harris, C., Arens, R.J., Srinivasan, P.: Trec blog and trec chem: A view from the corn fields. In: TREC. (2009)
[25] Santos, R.L.T., McCreadie, R.M.C., Macdonald, C., Ounis, I.: University of glasgow at trec 2010: Experiments with terrier in blog and web tracks. In: TREC. (2010)
[26] Yang, J., Dong, X., Guan, Y., Huang, C., Wang, S.: Hit_ltrc at trec 2010 blog track: Faceted blog distillation. In: TREC. (2010)
[27] Taboada, M., Grieve, J.: Analyzing appraisal automatically. In: Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications. (2004)
[28] Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phraselevel sentiment analysis. In: HLT-EMNLP'05. (2005)
[29] Bengtson, E., Roth, D.: Understanding the value of features for coreference resolution. In: EMNLP '08. (2008)
[30] Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: WWW '07. (2007)
[31] Marneffe, M.c.D., Maccartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: LREC'06. (2006)
[32] Fano, R.: Transmission of Information: A Statistical Theory of Communications. The MIT Press, Cambridge, MA (1961)
[33] Liu, S., Liu, F., Yu, C., Meng, W.: An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In: SIGIR '04. (2004)
[34] Macdonald, C., Ounis, I., Soboroff, I.: Overview of the trec 2010 blog track. In: TREC. (2010)
[35] Xu, X., Tan, S., Liu, Y., Cheng, X., Lin, Z., Guo, J.: Find me opinion sources in blogosphere: a unified framework for opinionated blog feed retrieval. In: WSDM '12. (2012)

LIFENG JIA
UNIVERSITY OF ILLINOIS AT CHICAGO,
CHICAGO, IL 60607, USA
E-MAIL: <LJIA2@UIC.EDU>

CLEMENT YU
UNIVERSITY OF ILLINOIS AT CHICAGO,
CHICAGO, IL 60607, USA
E-MAIL: <CYU@UIC.EDU>

WEIYI MENG
BINGHAMTON UNIVERSITY,
BINGHAMTON, NY 13902, USA
E-MAIL: <MENG@CS.BINGHAMTON.EDU>

LEI ZHANG
UNIVERSITY OF ILLINOIS AT CHICAGO,
CHICAGO, IL 60607, USA
E-MAIL: <LZHANG32@GMAIL.COM>