A Unified Relevance Model for Opinion Retrieval

Xuanjing Huang
School of Computer Science and Technology
Fudan University
Shanghai 200433, China
xjhuang@fudan.edu.cn

W. Bruce Croft
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003
croft@cs.umass.edu

ABSTRACT

Representing the information need is the greatest challenge for opinion retrieval. Typical queries for opinion retrieval are composed of either just content words, or content words with a small number of cue "opinion" words. Both are inadequate for retrieving opinionated documents. In this paper, we develop a general formal framework—the *opinion relevance model*—to represent an information need for opinion retrieval. We explore a series of methods to automatically identify the most appropriate opinion words for query expansion, including using query independent sentiment resources. We also propose a relevance feedback-based approach to extract opinion words. Both query-independent and query-dependent methods can also be integrated into a more effective mixture relevance model. Finally, opinion retrieval experiments are presented for the Blog06 and COAE08 text collections. The results show that, significant improvements can always be obtained by this opinion relevance model whether sentiment resources are available or not.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms

Algorithms, Experimentation, Theory

Keywords

Opinion retrieval, sentiment analysis, language model, opinion relevance model, relevance feedback, query expansion

1. INTRODUCTION

People have become increasingly interested in sharing their personal opinions and reviews about consumer products, commercial services, and even politics with others through online media. Their opinions can not only help other people to make decisions, but also help business and government agencies to collect valuable feedback. To support these activities, there is clearly a strong need for systems that can retrieve and analyze online opinions precisely and efficiently [21].

There has already been a considerable amount of research on sentiment analysis, including semantic orientation analysis of opinion words [24], sentiment classification of documents [20], subjectivity categorization of sentences [6], online opinion extraction [11] and opinion summarization [19]. In this paper, we will focus on opinion retrieval, which aims at automatically finding attitudes or opinions about specific targets, such as named entities, consumer products, or public events.

Opinion retrieval is very different to traditional topic-based retrieval. Firstly, relevant documents should not only be relevant to the targets, but also contain subjective opinions about them. Secondly, the text collections are more informal "word of mouth" web data. Two typical sources are blogs that generally reflect personal opinions and forums that present group opinions. Thirdly, although web retrieval pays more attention to precision, opinion retrieval attaches extra importance to recall, since further sentiment mining relies heavily on the coverage of the opinion collection.

Finally, the greatest challenge for opinion retrieval lies in the difficulty in representing the user's information need. For topic-based retrieval, the information need is usually defined by a user query consisting of a small number of keywords. Similarly, for opinion retrieval, typical queries are composed of either just content words, or content words as well as some cue words, for example, "*Steve jobs*", "*Find opinions about NASA*". Unfortunately, both of these are poor representation for opinion retrieval. It has been shown that if these types of queries are used directly, the average precision of opinion retrieval is much lower than that of topic-based retrieval [13]. In addition, only a small portion of relevant documents (about 25% in Blog06, a major text collection for opinion retrieval) contain cue words such as "opinion", "attitude", "review" and "sentiment". Simply submitting these words in queries to the retrieval system will hurt both recall and precision.

Currently, the majority of previous work in opinion retrieval treats this task as a two-stage process. In the first stage, documents are ranked by topical relevance only. In the second stage, candidate relevant documents are re-ranked by their opinion scores [16, 13]. The opinion scores can be acquired by either machine learning-based sentiment classifiers, such as SVM [29], external sentiment dictionaries with weighted scores from training documents [5, 1, 15], exhaustively computed query term–opinion word proximity scores [25, 31], or external toolkits such as *OpinionFinder* [5]. Although this two-stage process has been shown to be quite effective, the overall performance strongly depends on the initial topic-based retrieval and has high computational overheads, even if the opinion scores can be obtained during indexing [17].

In this paper, we will develop a general formal framework—

CIKM'09, November 2–6, 2009, Hong Kong, China.

the *opinion relevance model*—to directly represent the information need for opinion retrieval. According to this model, query terms are expanded with a small number of automatically chosen opinion words to represent the information need. We then explore a series of methods to extract the most appropriate sentiment words automatically: we will first study how to efficiently use query-independent sentiment resources such as opinionated seed words, annotated sentiment corpora, and relevance data to improve opinion retrieval; next we will propose a relevance feedback approach within the language modeling framework to combine the information about relevant or pseudo-relevant documents into the ranking algorithm. Both query-independent and query-dependent methods are also integrated into a more effective mixture relevance model.

The novelty and effectiveness of the proposed approach lie in: 1) the two-stage process of topic retrieval and sentiment classification can be converted to a unified opinion retrieval procedure through query expansion; 2) when relevance data are available, the improvement of this approach over the topic retrieval baseline is considerable and comparable to the best previous TREC results. When sentiment resources are not available, significant improvements can still be achieved; 3) this approach is not only suitable for English data, but also shown to be effective for a Chinese benchmark collection; 4) although this framework is proposed for opinion retrieval, it can be extended to other retrieval tasks where user queries are inadequate to express the information need, such as geographical information retrieval and music retrieval.

The rest of this paper is structured as follows. We first review related work in the next section. The formal framework of the opinion relevance model is presented in Section 3. After that, we propose a series of methods to automatically extract the most appropriate opinion words to expand the original queries. Section 5 describes the two benchmark collections, as well as some English and Chinese sentiment resources. Section 6 will present the experimental results in detail. The final part is the concluding remarks.

2. RELATED WORK

Related work can be found in three areas. The first is the recent developments in language model techniques in information retrieval, in particular, the relevance model. Our opinion relevance model is actually an extension of the relevance model for topic-based retrieval. The second includes the lexicon-based opinion finding techniques, which are used to get the opinion score of documents in two-stage opinion retrieval. Third, there has been some research on unified sentiment retrieval models.

2.1 Language models in information retrieval

The use of language models in information retrieval started with Ponte and Croft, who described a retrieval model based on multiple-Bernoulli language models [22]. In a language model, documents in a collection are viewed as models, and a query is regarded as a term sequence randomly sampled from these models. Then the documents are ranked by the probability that the query is sampled from the models of these documents.

Zhai and Lafferty described smoothing techniques for language modeling in information retrieval [9]. Among them, Dirichlet smoothing has proved to be very effective.

The basic language model approach represents the information need with query terms. Although it has a number of advantages, it is limited in terms of combining information about relevant documents into the ranking. Therefore, the *relevance model* was proposed to represent the topic covered by relevant documents, and then both queries and relevant documents are regarded as samples of text generated from the relevance model [10].

2.2 Lexicon-based opinion finding

Research on opinion retrieval has been significantly advanced by the TREC Blog track, where an opinion finding task was introduced to find public sentiment for given targets. TREC evaluations during last three years have shown that sentiment lexicon-based methods lead to good performance in two-stage opinion retrieval [17].

A lightweight lexicon-based statistical approach was proposed in [5]. In this method, the distribution of terms in relevant opinionated documents was compared to their distribution in relevant fact-based documents to calculate an opinion weight. These weights were used to compute opinion scores for every retrieved document. A weighted dictionary was generated from previous TREC relevance data in [1]. This dictionary was submitted as a query to a search engine to get an initial query-independent opinion score of all retrieved documents. Similarly, a pseudo opinionated word composed of all opinion words was first created, and then used to estimate the opinion score of a document in [15]. This method was shown to be very effective in TREC evaluations.

The query-independent sentiment expansion described here also requires an external lexicon. However, only the most frequent words are used for expansion, instead of all the opinion words in the lexicon. Therefore, heavy computational overheads are avoided.

2.3 Unified sentiment retrieval model

There has been some limited research on unified sentiment retrieval models.

Eguchi and Lavrenko proposed a sentiment retrieval model in the framework of generative language modeling [4]. They modeled a collection of natural language documents or statements, each of which consisted of some topic-bearing and some sentiment-bearing words. The sentiment was either represented by a group of predefined seed words, or extracted from a training sentiment corpus. This model was shown to be effective on the MPQA corpus.

Mei and Zhai tried to build a fine-grained opinion retrieval system for consumer products [14]. The opinion score for a product was a mixture of several facets. Due to the difficulty in associating sentiment with products and facets, the experiment was also tested in small scale text collections.

Zhang and Ye proposed a generative model to unify topic relevance and opinion generation [28]. This model led to satisfactory performance, but an intensive computation load was inevitable during retrieval, since for each possible candidate document, a opinion score was summed up from the generative probability of thousands of sentiment words.

Similar to these retrieval models, the proposed opinion relevance model also tries to unify topic and opinion relevance. Unlike them, our approach does not model document generation, but the information need instead, which is more straightforward and efficient.

3. FORMAL FRAMEWORK FOR OPINION RETRIEVAL

As we have mentioned, modeling the information need is the greatest challenge for opinion retrieval. In this section, we will extend the relevance model approach to incorporate the particular information needs required for opinion retrieval.

Suppose we can obtain an *opinion relevance model* from a query. Then we can compare this query directly with the documents and rank documents according to the *KL-divergence* between the two probability distributions of the opinion relevance model and document model [8]. Let R denote the opinion relevance model for a query, D denote the document model, and V denote the vocabulary,

then the KL-divergence between these two models is defined as:

$$KL(R||D) = \sum_{w \in V} P(w|R) log \frac{P(w|R)}{P(w|D)}$$
$$= \sum_{w \in V} P(w|R) log P(w|R) - \sum_{w \in V} P(w|R) log P(w|D)$$

$\sum_{w \in V} P(w|R) log P(w|R)$ is identical for all documents. Then documents can be scored by the reverse order of

$$Score(D) = \sum_{w \in V} P(w|R) log P(w|D)$$

$P(w|D)$ can be estimated effectively from the interpolation of the maximum likelihood estimation and the occurrence probability in the entire document collection, for example, *Dirichlet* smoothing [27]:

$$P(w|D) = \frac{freq_D(w) + \mu C_w/|C|}{|D| + \mu} \quad (1)$$

Where, $freq_D(w)$ is the occurrence frequency of w in D, C_w is the collection frequency of w, $|C|$ and $|D|$ are the number of word occurs in document D and collection C respectively, and μ is an empirical parameter (typically, 2,500).

Now the problem turns to the estimation for $P(w|R)$, the probability of word w given the model R. The basic language model simply estimates $P(w|R)$ from the occurrence of w within the query Q. Lavrenko and Croft's relevance model tries to estimate the probability of words from a group of documents relevant to the user query [10]. This model is proposed for topic-based retrieval. Here we will extend the relevance model to improve opinion retrieval, and call this new model the *opinion relevance model*.

Content words and opinion words contribute differently to opinion retrieval. Topical relevance is determined by the matching of content words and the user query, and the sentiment and subjectivity of a document is decided by the opinion words. So we divide the vocabulary V into two disjoint subsets: a content word vocabulary CV, and an opinion word vocabulary OV. Then we have:

Definition 1. An *opinion relevance model* is a unified model of both topic relevance and sentiment. In this model, $Score(D)$, the score of a document is defined as:

$$Score(D) = \alpha \sum_{w \in CV} P(w|R) log P(w|D) +$$
$$+ (1 - \alpha) \sum_{w \in OV} P(w|R) log P(w|D)$$

The parameter α is introduced to balance two relevance scores. The first part on the right hand side of the equation is just the same as the relevance model for topic-based retrieval. CV can be assigned as the set of original query terms, or obtained by any query expansion technique, which is outside the scope of this research. The key issue in our research is the selection of OV and the estimation of $P(w|R)$ for $w \in OV$, which is actually a special query expansion procedure. This procedure is called *sentiment expansion*, since only opinion words are used to expand an original query instead of content words.

4. SENTIMENT EXPANSION TECHNIQUES

Since the relevance scores should be computed as quickly as possible during retrieval, a smaller vocabulary is preferred. In the following subsections, we will explore several sentiment expansion methods to build the opinion word vocabulary with the most appropriate words. These methods can be divided into three categories. For query-independent sentiment expansion, we will make use of several kinds of sentiment resources. These resources include seed words, opinionated or general text corpora, and relevance data. For query-dependent sentiment expansion, we propose a relevance feedback-based approach. The query-independent and query-dependent methods can also be combined into a mixture model.

4.1 Query-independent sentiment expansion

4.1.1 Sentiment expansion based on seed words

This method restricts OV to some predefined seed words, such as those recommended by Turney [24]:

Positive seeds: good, nice, excellent, positive, fortunate, correct, superior.

Negative seeds: bad, nasty, poor, negative, unfortunate, wrong, inferior.

Seed words like "good" and "bad" are also adopted by Eguchi and Lavrenko's sentiment retrieval model [4].

In this case, the expanded query is composed of the original query and some additional opinionated seed words. The advantage of this method lies in that these seed words nearly always express strong sentiment. The disadvantage is that some of them are infrequent in the text collections.

4.1.2 Sentiment expansion based on text corpora

Instead of predefined seed words, we can obtain opinion words from lexical resources, such as *General Inquirer*[1], *OpinionFinder's Subjectivity Lexicon*[2].

There are always thousands of entries in a lexicon, so only the most frequent opinion words are selected to expand the original queries. That is to say, we can select opinion words according to their occurrence probabilities.

The occurrence probability can be estimated by the maximum likelihood method from any corpus, such as the text collection itself or the entire web. However, occurrences in a general corpus may be misleading. For example, "complete" is the second most frequent opinion word in the Blog06 collection, but it occurs more frequently in fact-based documents than in opinionated documents. In fact, it means "finish" or "finished" in most fact-based documents, and then is non-opinionated. Therefore, an opinionated corpus will be more reliable, such as the *Cornell movie review datasets* [20] and *MPQA Corpus* [26].

In the following experiments, the original queries are expanded with the most frequent opinion words in several opinionated or general corpora. We also find an appropriate number of opinion words in sentiment expansion.

4.1.3 Sentiment expansion based on relevance data

We now discuss the estimation of $P(w|R)$, or equivalently for ranking, the weight of w. For topic-based retrieval, a simple maximum likelihood estimate is often used in practice, based on the frequency in the query text ($freq_Q(w)$) and the number of words in the query($|Q|$) [3]:

$$P(w|R) \approx \frac{freq_Q(w)}{|Q|}$$

Since opinion words usually do not appear in the query text, this estimation is not applicable in the above methods. Therefore, the

[1]http://www.wjh.harvard.edu/~inquirer

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

probability is assumed to be uniform for the seed word or corpora-based sentiment expansion — that is to say, all opinion words are regarded as equally important.

Recently, "learning to rank" techniques have gained attention from both the information retrieval and machine learning communities. The goal is to automatically learn a function from training data to rank documents [7]. Query-independent features have been shown to be useful for ranking [2, 23].

Many opinion words, especially the most frequent opinion words, are also query-independent. For example, "good", "bad" can be used to modify almost any target, and "even", "too" can be used to modify almost any opinionated adjective. Therefore, we can make use of a simple machine learning technique to find the most valuable opinion words for sentiment expansion. Moreover, the weights of opinion words can also be obtained by learning.

Given a set of query relevance judgments, we can define the individual *contribution* to opinion retrieval for an opinion word:

Definition 2. If w is an opinion word, then the *contribution* of w means the maximum increase in the mean average precision (MAP) of the expanded queries over a set of original queries, where w is used to expand every original query.

More formally, let Q_i be the i-th query in the training set, $Q_i \bigcup w$, be the i-th expanded query, $weight(w)$ is the weight of w, $AP(Q_i)$ is the average precision of the retrieved documents for Q_i, and $AP(Q_i \cup w, weight(w))$ is the average precision of the retrieved documents for the expanded query while the weight of w is set as $weight(w)$, then,

$$Contribution(w) = \max_{weight(w)} \frac{1}{N} \sum_{i=1}^{N} \Big\{ AP(Q_i \cup w, weight(w)) - AP(Q_i) \Big\}$$

And also,

$$weight(w) = \arg\max \Big(Contribution(w) \Big)$$

The contribution of an opinion word can be used to assess to what extent it can improve the performance of opinion retrieval. After we learn the individual contribution for every opinion word, those words with the highest contribution will be used for sentiment expansion.

4.2 Query-dependent sentiment expansion

In the above methods, the selection of an opinion word is assumed to be independent of individual targets. On the other hand, a target is always associated with some particular opinion words. For example, "Mozart", the Austrian musician, is always regarded as a "genius" and "famous". Therefore, we can condition the probability of w on a query to incorporate the dependency between the target and the opinion word:

$$P(w|R) \approx P(w|Q) = P(w|q_1, q_2, \ldots, q_n)$$

Where, $q_1, q_2, \ldots, q_n$ are query terms in Q. In order to estimate the conditional probability, we propose a relevance feedback method to extract opinion words from a set of user-provided relevant opinionated documents. First, we have:

$$P(w|R) \approx P(w|q_1, q_2, \ldots, q_n) = \frac{P(w, q_1, q_2, \ldots, q_n)}{P(q_1, q_2, \ldots, q_n)}$$

$P(q_1, q_2, \ldots, q_n)$ can be obtained through marginalization:

$$P(q_1, q_2, \ldots, q_n) = \sum_{w} P(w, q_1, q_2, \ldots, q_n)$$

Then the joint probability of $P(w, q_1, q_2, \ldots, q_n)$ is estimated given the relevant document set of C:

$$P(w, q_1, q_2, \ldots, q_n) = \sum_{D \in C} P(D) P(w, q_1, q_2, \ldots, q_n | D)$$

The prior probability $P(D)$ is assumed to be uniform, while

$$P(w, q_1, q_2, \ldots, q_n | D) = P(w|D) P(q_1, q_2, \ldots, q_n | D, w)$$

$P(w|D)$ is also estimated by Dirichlet smoothing according to Equation (1). Assuming q_i is conditionally independent to q_j when both D and w are observed, then we have

$$P(q_1, q_2, \ldots, q_n | D, w) \approx \prod_{i=1}^{n} P(q_i | D, w)$$

$P(q_i|D, w)$ is then estimated by the co-occurrence of q_i and w within D:

$$P(q_i|D, w) = \begin{cases} freq_D(q_i)/|D| & \text{if } w \text{ occurs in } D \\ 0 & \text{otherwise} \end{cases}$$

The relevance feedback approach can still be used when there are no user-provided relevant documents. Under these circumstances, the relevant document set C can be acquired through pseudo-relevance feedback. We first rank documents using query likelihood scores, then select some top ranked documents to get the pseudo relevant set of C.

4.3 Mixture relevance model

We have proposed two types of sentiment expansion approaches, query-independent and query-dependent. For query-independent approaches, the most valuable opinion words are always general words and can be used to express opinions about any target. For query-dependent approaches, those words most likely to co-occur with the terms in the original query are used for expansion. These words are used to express opinions about particular targets. It is quite natural to integrate query-independent and query-dependent sentiment expansion into a *mixture relevance model* to cover both types of opinion words.

Definition 3. In a *mixture relevance model*, the final score of a document is defined as the interpolation of the scores assigned by original query, query-independent sentiment expansion, and query-dependent sentiment expansion:

$$\begin{aligned} Score(D) = \alpha \sum_{w \in Q} P(w|Q) log P(w|D) + \\ + \beta \sum_{w \in OV_1} P(w|R_1) log P(w|D) + \\ + (1 - \alpha - \beta) \sum_{w \in OV_2} P(w|R_2) log P(w|D) \end{aligned}$$

Where OV_1 and OV_2 are the sets of query-independent and query-dependent opinion words respectively, and $\alpha, \beta \in [0, 1]$.

5. TEST COLLECTIONS AND SENTIMENT RESOURCES

5.1 Benchmark data sets

The proposed methods will be verified on two benchmark collections, "Blog06" and "COAE08". Blog06 was created by the University of Glasgow for the blog retrieval track of TREC [12, 17]. This track has continued from 2006 to 2008, and 50 new queries are provided for evaluation every year. The 50 queries as well as

Table 1: Details for the Blog06 and COAE08 collections. The English translation of the Chinese topic is placed in parentheses.

Collection		Blog06	COAE08
Evaluation		TREC/Blog	COAE
Topic	Training	50	Not available
	Testing	100	20
	Example	*Mozart*	李连杰(*Jet Li*)
Documents	Number	3215K	40K
	Size	20GB	52M

the relevant opinionated documents in 2006 are used for training, while the other 100 queries are used for testing.

COAE08 is the benchmark data set of the opinion retrieval track of the first Chinese Opinion Analysis Evaluation (COAE), which was created by the Institute of Computing Technology, Chinese Academy of Sciences [30]. Since COAE has been held only once, training queries are not available.

Both evaluations aim at locating documents that express an opinion about a given target. The target can be not only a named entity, but also a concept, a product name, or an event. The relevance judgments were created by the pooling method, where documents are ranked at different levels: irrelevant, relevant but without opinion, and relevant with opinion. Table 1 shows some details for these two collections.

5.2 Sentiment resources

5.2.1 English resources

External sentiment lexicons provide the source of opinion words. Currently there are several online English sentiment lexicons: *General Inquirer* lists about 3,600 opinion words, and *OpinionFinder's Subjectivity Lexicon* lists more than 5,600 words.

Whether a word is opinionated or not is still debatable. For example, "home" and "just" are among the most frequent opinion words in *General Inquirer*, but are not listed in *OpinionFinder*, while "so" is listed in *OpinionFinder*, but not in *General Inquirer*.

Our opinion relevance model does not depend on the coverage of the sentiment lexicon, since an original query is expanded with only a small number of opinion words. In order to reduce the variability caused by different lexical resources, the intersection of these two lexicons is used instead.

Three English opinionated corpora are used for query-independent sentiment expansion: the *Cornell movie review datasets*, the *MPQA Corpus*, and the "Blog06(op)" opinionated data set, which is composed of the opinionated documents relevant to the 50 training queries.

Two general English collections are also used. One is the Blog06 collection itself. The other is "Web", which gives the Google hits for all opinion words.

Table 2 shows the most frequent 5 opinion words from these collections. From this table, we can see that this set varies with the corpus. Two opinionated corpora of "*Movie Reviews*" and "Blog06(op)" share a lot similar opinion words, while "*MPQA*"contains some distinct words. In fact, different to other corpora, the *MPQA* corpus belongs to the political genre and then contains a lot of formal opinion on political, economic, and governmental issues.

Table 3 shows some statistics about the most frequent opinion words from the "Blog06(op)" collection. "Average TF" is the average term frequency, "DF" is the document frequency, and "Coverage" is the percentage of documents in which the opinion word occurs. All of these opinion words occur several times in more than

Table 2: Most frequent 5 English opinion words from 3 opinionated and 2 general collections.

Corpus	Opinionated collection			General collection	
	Movie	MPQA	Blog06(op)	Blog06	Web
Documents	2000	535	11523	3.2M	14B
Most	like	against	like	like	free
frequent	even	minister	know	complete	back
English	good	terrorism	even	good	like
opinion	too	even	good	know	best
words	plot	like	too	free	show

Table 3: Statistics for the most frequent 5 opinion words in the "Blog06(op)" collection.

Word	Average TF	DF	Coverage
like	8.32	8200	71.2%
know	5.27	6970	60.5%
even	4.67	6521	56.6%
good	4.59	7047	61.2%
too	3.12	5998	52.1%

half of the opinionated documents. In fact, only about 10% opinionated documents contain none of them. This helps to explain why query-independent sentiment expansion can improve opinion retrieval significantly.

5.2.2 Chinese resources

For Chinese opinion retrieval, *HowNet Sentiment Vocabulary* is used as the sentiment lexicon[3]. It consists of about 7,000 opinion words. Two Chinese opinionated document sets are used in the experiments: "*Product*" is the data set of the opinion extraction track of the first COAE evaluation, which was created by Institute of Automation, Chinese Academy of Sciences and Fudan University [30]. This data set is composed of reviews of consumer products. The other is "*Hotel*", which is composed of reviews of hotels and was created by the Institute of Computing Technology[4]. Two general text corpora of COAE08 itself and the web are also provided. Since Google does not support Chinese segmentation, the web hits are provided by the *Sogou* Lab[5].

Table 4 shows the most frequent 5 opinion words from these collections. It can be found that almost all of them are single-character words with more than 10 ambiguous meanings (Sogou does not provide hits for single-character words). For example, "和" occurs frequently, and it has two opinionated meanings of "kind" and "peace", but it means "and" most of the time!

6. EXPERIMENTS

6.1 Performance of sentiment expansion approaches

Table 5 and 6 summarize the evaluation results for the Blog06 and COAE08 collections using our opinion relevance model.

The left columns in these tables show the sentiment expansion approaches. The results of the *baseline* system is first given, which is implemented with *Indri* search engine[6]. The baseline uses the basic relevance model as well as the Dirichlet smoothing technique

[3]http://www.keenage.com
[4]http://www.searchforum.org.cn/tansongbo/corpus-senti.htm
[5]http://www.sogou.com/labs/
[6]http://www.lemurproject.org

Table 4: Most frequent 5 Chinese opinion words from 2 opinionated and 2 general collections. English translations are in parentheses.

Corpus	Opinionated collection		General collection	
	Product	Hotel	COAE08	Sougo
Documents	478	4000	40K	1.5G
Most frequent Chinese opinion words	不(no)	不(no)	和(kind/peace)	可以(OK)
	有(possess)	很(much)	不(no)	没有(lack)
	也(also)	有(possess)	有(possess)	自己(of one's own side)
	和(kind/peace)	还(still)	也(also)	就是(quite right)
	就(exactly)	就(exactly)	中(all right)	不是(fault)

only. CV, the content word vocabulary, is assigned as the set of original query terms, and the document priors are set to be uniform.

All the sentiment expansion approaches are divided into three categories: query-independent, query-dependent and mixture relevance model. Each category is further divided into several subcategories, which will be explained in the following subsections.

The right columns show the evaluation results. The mean average precision (MAP) is the primary evaluation metric in both TREC Blog and COAE evaluations. Other metrics in these evaluations include R-precision (R-prec), binary Preference (bPref) and Precision at 10 documents (P@10).

6.1.1 Query-independent sentiment expansion

Query-independent sentiment expansion is further categorized by the sentiment resources, including the seed words, the opinionated corpus, the general corpus, and the relevance data. The seed words-based approach is not applicable for the COAE08 collection, since there are no generally accepted Chinese seed words. Relevance data-based sentiment expansion is also not available for COAE08. There are two runs based on the seed words for the Blog06 collection. In "Seed-1", the final query is represented with the original query as well as a single pair of seed words of "good" and "bad". In "Seed-7", seven pairs of seed words are given, which are the same as [24]. The top 5 opinion words are used for sentiment expansion in other query-independent runs. Here, all corpus-based query-independent runs are named by the associated corpus, and "RD" is the relevance data-based run.

From these tables we can see that the seed word approach is not always helpful. In fact, its effectiveness is dependent on the selection of seed words. Although these seed words are typical opinion words with strong and unambiguous sentiment, only a small portion of them frequently appear in the opinionated documents. Among the most frequent 50 English opinion words, only "good", "bad", "nice" and "poor" are chosen as seed words.

Previous studies show that sentiment dictionary-based methods lead to good performance in two-stage opinion retrieval, especially when statistical information obtained from relevance data is available [17]. Our experiments also verify this. When such a corpus is not available, other opinionated corpora are also helpful: significant improvement over the baseline approach can be achieved using the "*Movie Review*" and the "Hotel" corpus. Here, the Wilcoxon signed-rank test is used to test the differences between runs at a significance level of 0.05. If the annotated corpus is absent, marginal improvements may still be achieved with the help from a general text collection. When the most frequent opinion words in the Blog06 collection are used to expand the original queries, the improvement of MAP is still significant.

On the other hand, corpus based-sentiment expansion is sensitive to the resources. For example, when MPQA is used, MAP decreases. In fact, this corpus is very different to the Blog06 collection. A similar phenomenon happens in the "Product" run.

Table 7: Highest-contribution opinion words for the Blog06 and COAE08 collection.

Blog06		COAE08	
Word	Contribution	Word	Contribution
even	7.9%	不(no)	2.6%
like	7.7%	对(right)	1.4%
know	4.4%	很(much)	1.2%
too	4.2%	能(able)	1.2%
good	4.0%	就(exactly)	1.1%

Given the relevance data for 50 queries from 2006, we can estimate the individual *contribution* of all English opinion words for the Blog06 collection. We first calculate the MAP using the original queries. Then each opinion word is used to expand the original queries with a group of predefined weights, and the MAP using these expanded queries is also obtained. The most significant improvement is assigned as the contribution. The contributions of Chinese words for the COAE08 collection are also estimated from the overall relevance judgments just for comparison. Those opinion words that contribute the most to sentiment expansion are shown in Table 7.

It is interesting to notice that those 5 English opinion words with the highest contribution are the most frequent words in the Blog06(op) corpus but in a different order. Because the machine learning approach can assign weights for the opinion words more accurately, RD performs significantly better than Blog06(op).

We can also find that the contributions of Chinese words are much lower than those of English words. This is the major reason why the performance improvement on the COAE08 collection is not as significant as the Blog06 collection. Other reasons include the semantic ambiguity of Chinese opinion words, and the lack of a training corpus similar to COAE08.

6.1.2 Query-dependent sentiment expansion and mixture model

The query-dependent approach is based on pseudo relevance feedback. For PRB, the pseudo relevance feedback run, 20 query-dependent opinion words were extracted from 5 top-ranked documents. The mixture models are combined with PRB and the best performing query-independent runs using three different sentiment resources, which are Blog06(op), Blog06 and RD.

Table 5 and 6 show that pseudo relevance feedback does significantly improve opinion retrieval. Table 8 gives some examples of the opinion words with the highest conditional probability given the original queries. Some of them are still general terms, but a lot of opinion words are now strongly associated with the original queries — we can extract "genius" and "famous" for the musician "Mozart", as well as "protective" and "fidelity" for the organization "Allianz".

The mixture relevance model effectively integrates the query-

Table 5: Comparison of opinion-finding MAP, R-prec, bPref, P@10 for different sentiment expansion approaches for the Blog06 collection. The original queries are extracted from the title field of the topics automatically. The best in each column is highlighted.

Approach			Evaluation metric			
Category	Sub-category	Run id	MAP	R-prec	bPref	P@10
Baseline	/	Baseline	0.2655	0.3252	0.2974	0.4770
Query-independent	Seed words	Seed-1	0.2797	0.3335	0.3120	0.5250
		Seed-7	0.2650	0.325	0.3058	0.4690
	Opinionated corpus	Movie	0.2961	0.3422	0.3303	0.5460
		MPQA	0.2732	0.3315	0.3082	0.4880
		Blog06(op)	0.3097	0.3530	0.3395	0.5570
	General corpus	Blog06	0.2822	0.3340	0.3133	0.5200
		Web	0.2733	0.3313	0.3055	0.5100
	Relevance data	RD	0.3117	0.3542	0.3408	0.5650
Query-dependent	Pseudo relevance feedback	PRB	0.2806	0.3333	0.3101	0.4950
Mixture model	Blog06(op)+PRB	MBoP	0.3124	0.3521	0.3404	**0.5670**
	Blog06+PRB	MBP	0.3009	0.3477	0.3340	0.5480
	RD+PRB	MRDP	**0.3147**	**0.3546**	**0.3418**	0.5640

Table 6: Comparison of opinion-finding MAP, R-prec, bPref, P@10 for different sentiment expansion approaches for the COAE08 collection. The best in each column is highlighted.

Approach			Evaluation metric			
Category	Sub-category	Run id	MAP	R-prec	bPref	P@10
Baseline	/	Baseline	0.3565	0.4046	0.3874	0.7700
Query-independent	Opinionated corpus	Product	0.3597	0.4149	0.3932	0.7750
		Hotel	0.3658	0.4240	0.4011	0.7700
	General corpus	COAE08	0.3621	0.4174	0.3959	0.7550
		Sougo	0.3571	0.4139	0.3880	0.7700
Query-dependent	Pseudo relevance feedback	PRB	0.3677	0.4273	0.4031	0.7600
Mixture model	Hotel+PRB	MHP	**0.3697**	**0.4311**	**0.4069**	0.7750
	COAE08+PRB	MCP	0.3685	0.4286	0.4060	**0.7900**

Table 8: Highest-probability opinion words for 6 example queries in the Blog06 and COAE08 collections. The probabilities are estimated by the PRB run from the top 5 documents.

Blog06			COAE08		
Mozart	Allianz	Wikipedia	李连杰(Jet Li)	宏观调控(Macro-regulation)	CPI(Consumer Price Index)
like	best	like	不(no)	就(exactly)	不(no)
good	premium	open	有(possess)	还(still)	有(possess)
too	great	good	而(further)	改革(reform)	也(also)
even	value	know	但(but)	认为(believe)	上涨(rise up)
death	traditional	free	多(much)	继续(continue)	而(further)
best	independent	great	精彩(wonderful)	有(possess)	都(all)
great	protective	excellent	认为(believe)	问题(problem)	就(exactly)
genius	unique	best	能(able)	进一步(further)	继续(continue)
famous	fidelity	knowledge	却(but)	多(much)	问题(problem)
favorite	alliance	authoritative	很(much)	高(high)	压力(pressure)

Figure 1: Results of varying the number of opinion words for the Blog06 collections using both query-independent and query-dependent sentiment expansion. Three typical query-independent runs of Blog06(op), Blog06, and RD, as well as the query-dependent runs of PRB are compared. The X-axis shows the number of opinion words used in sentiment expansion. The Y-axis gives the corresponding MAP.

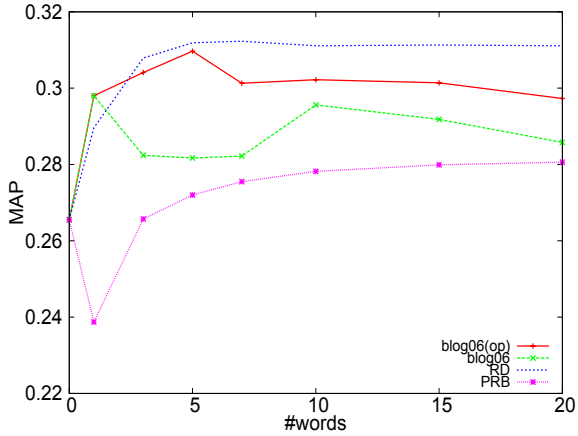
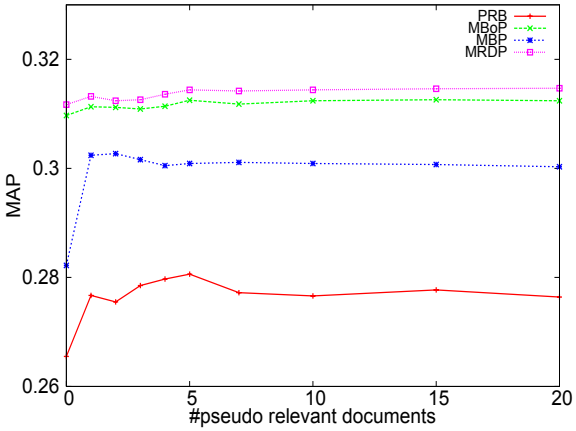


Figure 2: Results of varying the number of pseudo relevance documents for the Blog06 collections using query-dependent sentiment expansion and mixture relevance model. The query-dependent runs of PRB, as well as all three mixture runs MBoP, MBP, and MRDP are compared. The X-axis shows the number of pseudo relevance documents. The Y-axis gives the corresponding MAP.



independent and query-dependent approaches. The mean average precision of each mixture run is significantly better than both components for the Blog06 collection.

TREC Blog evaluations show that the most significant improvement on MAP over the the topic-relevance baseline is 17.0% for 50 queries from 2007, and the best average opinion-finding improvements over the standard topic baselines is 11.8% for 50 queries from 2008 [13, 18]. It can be observed from Table 5 that the improvement on MAP over the baseline reaches as high as 18.5% for our best run of MRDP. The improvement on MAP for Blog06(op), RD, and MBoP are also higher than 16.0%. For the COAE08 collection, the increase in MAP for MHP is also a little higher that those of "Hotel" and PRB, although the improvement is not as significant.

6.2 Discussion

We also investigated whether more opinion words lead to better retrieval performance. Figure 1 shows the change of MAP when the number of opinion words varies for the Blog06 collections.

We find that MAP is improved as soon as opinion words are combined into the model in query-independent runs. Why does sentiment expansion with such a small number of opinion words lead to such promising results? The reason is due to the wide coverage of these words. However, MAP is not improved further when more than 10 terms are used for expansion. The reason is probably due to the lack of specificity in query-independent methods.

For the query-dependent run of PRB, satisfactory improvement can be achieved when 5~10 opinion words are chosen for expansion. When more opinion words are selected, performance still increases slowly.

Considering these factors, 5 query-independent and 20 query-dependent opinion words are used for sentiment expansion.

For query-dependent sentiment expansion, another important factor is the number of pseudo relevance documents. Figure 2 shows the change of MAP when the number of pseudo relevance documents varies for the Blog06 collection. It shows that MAP increases significantly even only 3~4 documents are employed for pseudo relevance feedback, and then remains stable or fluctuates somewhat. The reason is that not all documents with high ranking scores are relevant and opinionated, and the possibility of this decreases as the scores decrease. Therefore, 5 documents are employed for pseudo relevance feedback.

Table 9: Caparison of topic-relevance MAP, R-prec, bPref and P@10 for the Blog06 collection. The best in each column is highlighted.

Run	MAP	R-prec	bPref	P@10
Baseline	0.3335	0.3866	0.3891	0.6190
Blog06(op)	0.3691	0.4092	0.4241	0.6690
Blog06	0.3450	0.3923	0.4019	0.6540
RD	0.3688	**0.4099**	**0.4268**	**0.6780**
PRB	0.3429	0.3860	0.3949	0.6230
MBoP	0.3681	0.4070	0.4249	0.6770
MBP	0.3577	0.4042	0.4162	0.6550
MRDP	**0.3702**	0.4080	0.4264	0.6720

We also observe the effects on retrieval effectiveness when varying the parameters of α and β in Definitions 1 and 3, which are employed to adjust the weight of query-independent and query-dependent sentiment expansion approaches. Figure 3 gives the MAP surface of the run MRDP for the Blog06 collection. Surfaces of other mixture runs are similar. We have found that the surfaces are always concave or very close to concave, and they always have the same general form. Therefore, a simple hill-climbing search can be used to optimize MAP. Since the surface is almost concave we are likely to find the global maximum. For example, the optimal parameters for this run are $\alpha = 0.4$ and $\beta = 0.4$.

TREC evaluations show that a strongly performing topic-based retrieval baseline is very important in achieving good opinion finding retrieval performance [17]. We could also ask whether good opinion retrieval will improve topic-based retrieval. Table 9 shows the evaluation results for the topic-based retrieval for the Blog06 collection.

It can be observed that all these runs improve topic-based re-

Figure 3: MAP surface over simplex of parameter values using MRDP for the Blog06 collection. α ranges from 0.1 to 1.0, and β ranges from 0.0 to $1.0 - \alpha$. The step size is 0.1. Z-axis shows the MAP.

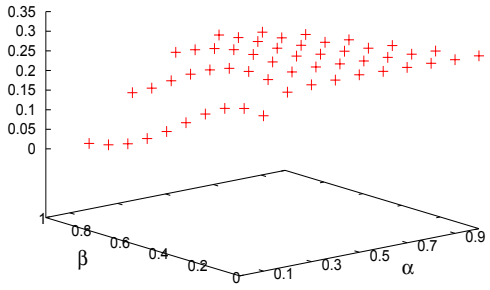
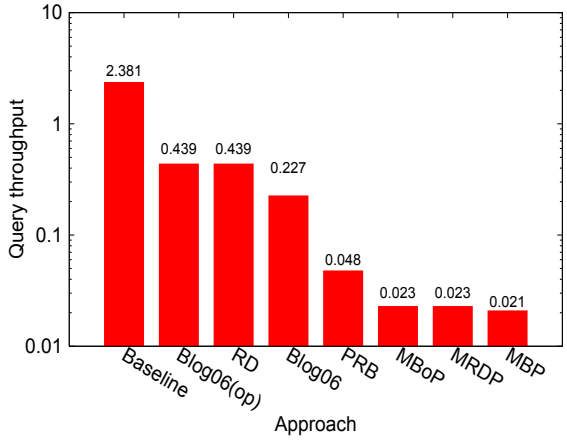


Figure 4: Query throughput of different sentiment expansion approaches on an Intel Xeon 3.00 GHz node for the Blog06 collection.



trieval, and the improvements are significant except PRB. This means that the sentiment expansion approaches (in particular, query-independent approaches) are not only effective for opinion retrieval, but also for topic-based retrieval for such targets as named entities, products or concepts. The reason is because these targets are often reviewed by web users.

Efficiency is another important issue in information retrieval. Figure 4 shows the query throughput (that is, the number of queries processed per second) of some sentiment expansion approaches. Obviously, the baseline approach leads to the highest throughput. It can also be observed that query-independent sentiment expansion are much faster than query-dependent approaches. More expansion terms lead to more processing time. It also takes some time to extract opinion words from the pseudo relevant documents. Therefore, query-dependent and mixture approaches result in lower throughput. However, these approaches are still much faster than two-stage opinion retrieval, since only top-ranked documents are considered instead of all the retrieved documents. Considering both effectiveness and efficiency factors, we can conclude that:

1. If retrieval effectiveness is preferred, mixture approaches should be adopted, since the combination of query-independent and dependent sentiment expansion leads to better retrieval performance. In our experiment, when query relevance data are available, the highest MAP can be achieved in MRDP.

2. If retrieval efficiency is preferred, query independent sentiment expansion should be adopted. When relevance data are available, the performance is only somewhat worse than the best mixture approach.[7] When the annotated data is absent, we can still choose the most frequent opinion words in the text collection to expand the original query.

7. CONCLUSION

In this paper, we have proposed the *opinion relevance model*, a formal framework for directly modeling the information need for opinion retrieval. In this framework, query terms are expanded with a small number of opinion words to represent the information need. We then propose a series of sentiment expansion approaches to find the most appropriate query-independent or query-dependent opinion words.

The proposed model has been verified on the Blog06 and COAE08 collections. The results show that very significant improvements can be obtained. We have also discussed the factors in opinion retrieval, including the number of opinion words in the expanded query, the number of documents in pseudo relevance feedback, the parameters in mixture relevance model, the impact of opinion retrieval on topic-based retrieval, as well as the efficiency issues.

Currently, the pseudo relevance feedback documents are ranked simply by their generative probabilities from the relevance model. As future work, we will take into consideration the diversity of the feedback documents, in order to retrieve more information about different facets of queried targets. Another line of interest is varying the document priors. In our opinion relevance model, the document priors are set to be uniform. In fact, blogs and forum are widely used to express opinions, and their layout, link structure and user behavior may also be helpful to judge the quality and popularity of opinions. We plan to incorporate this information as a document prior into the mixture relevance model.

8. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the National Natural Science Foundation of China (Grant NO. 60673038), in part by the Ph.D Programs Foundation of Ministry of Education of China (Grant NO. 200802460066), and in part by the Shanghai Committee of Science and Technology, China (Grant No. 08511500302). Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor. We thank Wang Bingqing, Xu Hongbo and Qian Xian for providing the Chinese sentiment resources. We also thank David Fisher for his helpful advice and support during algorithm implementation.

9. REFERENCES

[1] G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi. FUB, IASI-CNR and University of Tor Vergata

[7]We should notice that, the computation of the individual contribution of opinion words also takes some time. In fact, it takes about 120 seconds to calculate the contribution of an opinion word on the 50 training queries using an Intel Xeon 3.00 GHz node. However, the computation is implemented during training and will not slow the retrieval speed.

at TREC 2007 Blog Track. In *Proceedings of the 15th Text REtrieval Conference (TREC 2007)*, 2007.

[2] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML '05: Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96, New York, NY, USA, 2005. ACM.

[3] W. B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison Wesley, 1 edition, February 2009.

[4] K. Eguchi and V. Lavrenko. Sentiment retrieval using generative models. In *EMNLP '06, Proceedings of 2006 Conference on Empirical Methods in Natural Language Processing*, pages 345–354, 2006.

[5] D. Hannah, C. Macdonald, J. Peng, B. He, and I. Ounis. University of Glasgow at TREC 2007: Experiments in Blog and Enterprise Tracks with Terrier. In *Proceedings of the 15th Text REtrieval Conference (TREC 2007)*, 2007.

[6] V. Hatzivassiloglou and J. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, 2000.

[7] T. Joachims, H. Li, T.-Y. Liu, and C. Zhai. Learning to rank for information retrieval (LR4IR 2007). *SIGIR Forum*, 41(2):58–62, 2007.

[8] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[9] J. D. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 111–119, 2001.

[10] V. Lavrenko and W. B. Croft. Relevance-based language models. In *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 120–127, 2001.

[11] B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th International Conference on World Wide Web*, 2005.

[12] C. Macdonald and I. Ounis. The TREC Blogs06 collection: creating and analysing a blog test collection. Technical Report TR-2006-224, Department of Computer Science, University of Glasgow, 2006.

[13] C. Macdonald and I. Ounis. Overview of the TREC-2007 Blog Track. In *Proceedings of the 15th Text REtrieval Conference (TREC 2007)*, 2007.

[14] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, 2007.

[15] S.-H. Na, Y. Lee, S.-H. Nam, and J.-H. Lee. Improving opinion retrieval based on query-specific sentiment lexicon. In *ECIR '09: Proceedings of the 31st annual European Conference on Information Retrieval*, pages 734–738, 2009.

[16] D. Oard, T. Elsayed, J. Wang, Y. Wu, P. Zhang, E. Abels, J. Lin, and D. Soergel. TREC-2006 at Maryland: Blog, Enterprise, Legal and QA Tracks. In *Proceedings of the 15th Text REtrieval Conference (TREC 2006)*, 2006.

[17] I. Ounis, C. Macdonald, and I. Soboroff. On the TREC Blog Track. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2008.

[18] I. Ounis, C. Macdonald, and I. Soboroff. Overview of the TREC-2008 Blog Track. In *Proceedings of the 16th Text REtrieval Conference (TREC 2008)*, 2008.

[19] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 271–278, 2004.

[20] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124, 2005.

[21] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[22] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 275–281, 1998.

[23] M. Taylor, H. Zaragoza, N. Craswell, S. Robertson, and C. Burges. Optimisation methods for ranking functions with multiple parameters. In *CIKM '06: Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 585–593, New York, NY, USA, 2006. ACM.

[24] P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[25] O. Vechtomova. Using subjective adjectives in opinion retrieval from blogs. In *Proceedings of the 15th Text REtrieval Conference (TREC 2007)*, 2007.

[26] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation (formerly Computers and the Humanities)*, 39(2/3):164–210, 2005.

[27] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214, 2004.

[28] M. Zhang and X. Ye. A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In *SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 411–418, New York, NY, USA, 2008. ACM.

[29] W. Zhang and C. Yu. UIC at TREC 2007 Blog Track. In *Proceedings of the 15th Text REtrieval Conference (TREC 2007)*, 2007.

[30] J. Zhao, H. Xu, X. Huang, S. Tan, K. Liu, and Q. Zhang. Overview of Chinese Opinion Analysis Evaluation 2008. In *Proceedings of the First Chinese Opinion Analysis Evaluation*, 2008.

[31] G. Zhou, H. Joshi, and C. Bayrak. Topic categorization for relevancy and opinion detection. In *Proceedings of the 15th Text REtrieval Conference (TREC 2007)*, 2007.