Fusion Approach to Targeted Opinion Detection in Blogosphere

블로고스피어에서 주제에 관한 의견을 찾는 융합적 의견탐지방법

Kiduk Yang*

〈Contents〉

ABSTRACT

This paper presents a fusion approach to sentiment detection that combines multiple sources of evidence to retrieve blogs that contain opinions on a specific topic. Our approach to finding opinionated blogs on topic consists of first applying traditional information retrieval methods to retrieve blogs on a given topic and then boosting the ranks of opinionated blogs based on the opinion scores computed by multiple sentiment detection methods. Our sentiment detection strategy, whose central idea is to rely on a variety of complementary evidences rather than trying to optimize the utilization of a single source of evidence, includes High Frequency module, which identifies opinions based on the frequency of opinion terms (i.e., terms that occur frequently in opinionated documents), Low Frequency module, which makes use of uncommon/rare terms (e.g., "sooo good") that express strong sentiments, IU Module, which leverages n-grams with IU (I and you) anchor terms (e.g., I believe, You will love), Wilson's lexicon module, which uses a collection-independent opinion lexicon constructed from Wilson's subjectivity terms, and Opinion Acronym module, which utilizes a small set of opinion acronyms (e.g., imho). The results of our study show that combining multiple sources of opinion evidence is an effective method for improving opinion detection performance.

Keywords: Fusion, Information retrieval, Opinion detection

초 록

이 논문은 여러가지 자료를 결합해 어떤 주제에 관한 의견이 실려있는 블로그를 찾는 융합적 의견탐지방법을 소개한다. 주제에 관한 의견이 담긴 블로그를 찾기위해 이 연구는 기존의 IR 방법으로 주제에 관한 블로그를 검색한 후 여러가지 의견탐지 방법을 합산한 의견점수로 검색결과의 순위를 조정하는 방법을 쓴다. 의견탐지 모듈의 주요 구성 요소는 의견이 실려있는 블로그에 자주 나오는 단어들을 활용한 고빈도 모듈, 강한 감정을 표현하는 희귀 한 용어들을 (e.g., "sooo good") 활용한 저빈도 모듈, "I"와 "you"에 묶인 n-gram을 (e.g., I believe, You will love) 활용한 IU모듈, 윌슨의 주관 용어 목록을 바탕으로 한 윌슨의 어휘 모듈, 그리고 소수의 의견 약어를 (e.g., imho) 이용한 의견 약어 모듈들 이다. 본 연구의 결과는 여러 가지 방법을 융합하는 것이 의견 검출 성능을 향상시키는데 효과적이 다는 것을 보여주었다.

키워드: 융합, 정보 검색, 의견 감지

* Associate Professor, Department of Library and Information Science, Kyungpook National University (kiyang@knu.ac.kr)

· 논문접수: 2014년 11월 20일 ·최초심사: 2014년 11월 25일 ·게재확정: 2015년 3월 19일
· 한국도서관·정보학회지 46(1), 321-344, 2015. [http://dx.doi.org/10.16981/kliss.46.201503.321]

Ⅰ. Introduction

Blogs (short for Web logs), which are journal-like Web pages that started out as online diaries over a decade ago, have evolved in recent years to become one of the mainstream tools for collaborative content creation on the Web. Among the many characteristics of the blog, perhaps the most distinguishing is its highly personalized nature, often containing personal feelings, perspectives, and opinions about a topic. The blogosphere, with its rich collection of commentaries, is an important source for public opinion. However, finding "opinionated" blogs on a specific target (e.g., product, event, etc.) is a difficult task with compound challenges of Web search and opinion mining.

Even if topically relevant blogs were to be retrieved, identifying opinionated posts among them can be quite challenging due to the context-dependent nature of the subjective language. According to Wiebe et al. (2004), "both opinionated and factual documents tend to be composed of a mixture of subjective and objective language." In other words, it is hard to differentiate opinionated documents from factual ones with simple clues. Furthermore, the content characteristics of the blog, namely the high level of noise (e.g., advertisements) and prevalent use of non-standard language (e.g., word morphing), exacerbate the already difficult problem of combining Web information retrieval (IR) and sentiment detection approaches.

This paper presents our approach to finding opinionated blogs, which consists of first applying traditional information retrieval (IR) methods to retrieve blogs about a given topic and then boosting the ranks of opinionated blogs based on the opinion scores generated by multiple assessment methods. The central idea underlying our opinion detection method, which is to rely on a variety of complementary evidences rather than trying to optimize a single approach, is motivated by our past experience that suggested that neither the lexicon-based approach nor the machine learning approach is well-suited for the blog opinion retrieval task (Yang et al. 2006; 2007a).

Ⅱ. Prior Research

1. Opinion Detection

The first group of studies we review comes from a body of work that explores methods of detecting opinions or sentiments in documents. Among the rich spectrum of literature in this area, we select a few that are relevant to opinion mining in blogosphere.

Wiebe et al. (2004) and Wilson et al. (2003), who introduced practical methods for learning the subjective language from text corpora using information extraction and text categorization, established the theoretical grounding in subjectivity recognition. They presented several categories of subjectivity clues, which included low-frequency words, n-gram collocations, and adjectives and verbs, and evaluate their distributional similarities for performing the subjectivity recognition tasks. Although identifying distributional similarities requires a document collection to be pre-analyzed, it does not require document labels (i.e., “subjective” or “objective” annotations). Therefore, these approaches can be implemented without training data.

By using text analysis and external knowledge (i.e. Amazon’s Web Services for locating products), Mishne and de Rijke (2006) presented a method for analyzing blogs to derive product wish lists. Specifically, they suggested an approach to identify references to books based on keyword proximity (e.g., “read” and “book”) and relevant patterns (e.g., [ENTITY] by [PERSON]). They extracted keywords that appear frequently on a target set of blogs (e.g., book reviews) and identified words and patterns that are used for commenting purposes (e.g., “try”, “product”, “released by [ENTITY]”), etc).

Following a similar research direction, Liu, Hu, and Cheng (2005) proposed a framework for analyzing and comparing customer reviews of products. Employing a language pattern mining technique to extract product features, Liu et al. implemented a system called Opinion Server that integrates visualization methods to present retrieved results and compare customer reviews. Hu and Liu (2004) examined the problem of generating feature-based summaries of product reviews. Given a set of reviews of a particular product, authors divided the process of generating summaries into three subtasks: identifying sentences with customer opinions on products, identifying product features with customer opinions, and producing a summary of product reviews (i.e., features and opinions). Hui and Liu used association mining to find frequent noun phrases as product features and utilized WordNet to predict the semantic orientations of adjectives as opinion words. Instead of classifying the review as a whole, they

employed sentence-level classification to identify the orientation (i.e., negative or positive) of each sentence. At a later study, Liu (2012) investigated identification of “implicit sentiment” via a two-phase co-occurrence association rule that match implicit expressions of sentiment (e.g., “too heavy”) with explicit aspects (e.g., “weight”).

Chklovski (2006), who focused on automatic summarization of opinions and assessments in product reviews, discussion forums, and blogs, presented a system called GrainPile, which recognizes subjective expressions (e.g., fairly, very, extremely) and maps them to a common scale. Results from the study showed that this approach of quantifying the degree of opinions by subjective adverbs and adverbial phrases (e.g., “fairly”, “very”, “not too”, “pretty darn”, etc.) strongly outperformed the simple co-occurrence based method.

Others investigated the lexicon-based sentiment detection method, where a dictionary of sentiment words and phrases with associated strength and orientation of sentiment is used to compute a sentiment score of documents (Ding, Liu and Yu, 2008; Taboada et al. 2011). Studies comparing the lexicon-based approach with machine learning approach to sentiment detection found similar levels of performance (Thelwall et al. 2010; 2012)

2. TREC Blog Track

The content characteristics of blogs, namely the high level of non-posting content generated by blogware or spamming and non-standard usage of language often imbued with the informal tone of bloggers, pose new challenges for conventional opinion detection strategies that make use of syntactic cues in formal discourses or standard text analysis leveraging high frequency patterns.

Text Retrieval Conference (TREC), an international research forum that supports a variety of cutting-edge IR research, initiated the exploration of blog opinion mining in one of its specialized venues called the Blog Track in 2006 (Blog06), where methods that "uncover the public sentiment towards a given entity/target" in blogosphere were investigated. In 2009, the Blog Track introduced new search tasks of faceted blog distillation and top story identification with a new test collection (Blog08) but the revamped block track was soon replaced with Microblog Track that deals with searching microblogs such as Twitter in 2011 (Macdonald et al. 2010; Ounis et al. 2011). Since the focus of the study is on exploring targeted sentiment detection in blogosphere, the review of prior research in this section is limited to the original

Blog Track task of finding opinionated blogs on a given topic.

To find opinionated blogs on a given topic, most TREC participants employed a two-stage approach that involves an initial step of retrieving topically relevant blogs followed by a re-ranking step leveraging opinion finding features (Ounis et al. 2007). Opinion finding strategies in TREC generally fall into either classification-based or lexicon-based category. The classification-based approaches (Joshi, Bayrak and Xu 2007; Zhang and Zhang 2007; Zhang and Yu 2007) use training data, typically product or movie reviews for positive and news or encyclopedia documents for negative training data, to train a text classifier for identifying opinionated blogs, whereas the lexicon-based approaches construct a dictionary or lexicon of opinion terms and use the frequency of opinion terms in documents to compute the opinion scores (Mishne 2007, Oard et al. 2007; Yang et al. 2007a; Yang 2009).

Ⅲ. Research Question

Having explored the topical search problem over the years (Yang 2002; Yang and Albertson 2004; Yang et al. 2005), we focused on the question of how to adapt a topical retrieval system for opinion detection task. The intuitive answer was to first apply IR methods to retrieve blogs about a target (i.e., on-topic retrieval), and then identify opinionated blogs by leveraging evidences of opinion. Consequently, the key research question of our study concerns the evidences of opinion, namely what they are and how they can be leveraged. To maximize the total coverage of opinion evidence, we considered the following three complementary sources of evidence:

- Opinion Lexicon: One obvious source of opinion is a set of terms often used in expressing opinions (e.g., "Skype *sucks*", "Skype *rocks*", "Skype is *cool*").
- Opinion Collocations: One of the contextual evidence of opinion comes from collocations used to mark adjacent statements as opinions (e.g., "*I believe* God exists", "God is dead *to me*").
- Opinion Morphology: When expressing strong opinions or perspectives, people often use morphed word form for emphasis ("Skype is *soooo* buggy", "Skype is *bugfested*").

Because blogs are generated by content management software (i.e. blogware), which allows authors to create and update contents via a browser-based interface, they are laden with

non-posting content for navigation, advertisement, and formatting display. Therefore, the question of how such blogware-generated noise influences opinion detection merits consideration. In investigating the blog noise question, however, we did not include spam detection so as to determine whether the high noise level typical of blogs has adverse influence on opinion detection performance.

Ⅳ. Methodology

Our topical retrieval strategy in the blog track was driven by the hypothesis that query should be broad to ensure high recall. This strategy, which was realized in the form of minimal query processing without any query expansion, turned out to be a double-edged sword that worked well for longer queries that include more complete description of information need but performed poorly with short queries. To address the weakness of our past topical retrieval strategy, we extend our initial hypothesis as follows: “query should be broad yet descriptive enough to ensure high recall.” In accordance with the amended hypothesis regarding topical retrieval, the follow-up study reported in the paper investigated various Web-based query expansion methods to increase the initial retrieval performance for the short query.

Our opinion detection strategy is shaped by the hypothesis that there are multiple sources of opinion evidence that complement one another. Conventional opinion detection methods look to the most obvious source of the opinion evidence, namely the set of terms commonly used to express opinions (e.g., “Skype sucks”, “Skype rocks”, “Skype is cool”). The standard opinion terminology, or Opinion Lexicon as we call it, occurs frequently in opinionated documents while occurring infrequently in non-opinionated documents. Another source of opinion evidence that lies at the opposite end of the spectrum is what we call Opinion Morphology, which consists of morphed word forms used to express strong opinions or perspectives. The opinion morphologies, taking the form of compound word (e.g., “Skype is metacool”), repeated characters (e.g., “Skype is soooo buggy”), or intentional misspellings (e.g., “I luv Skype”), are creative expressions with low occurrence frequencies, which were found to be useful clues for sentiment detection (Wiebe et al. 2004). Opinion Collocation, phrases with a personal pronoun anchor (e.g, I, you) that mark adjacent statement as opinions,

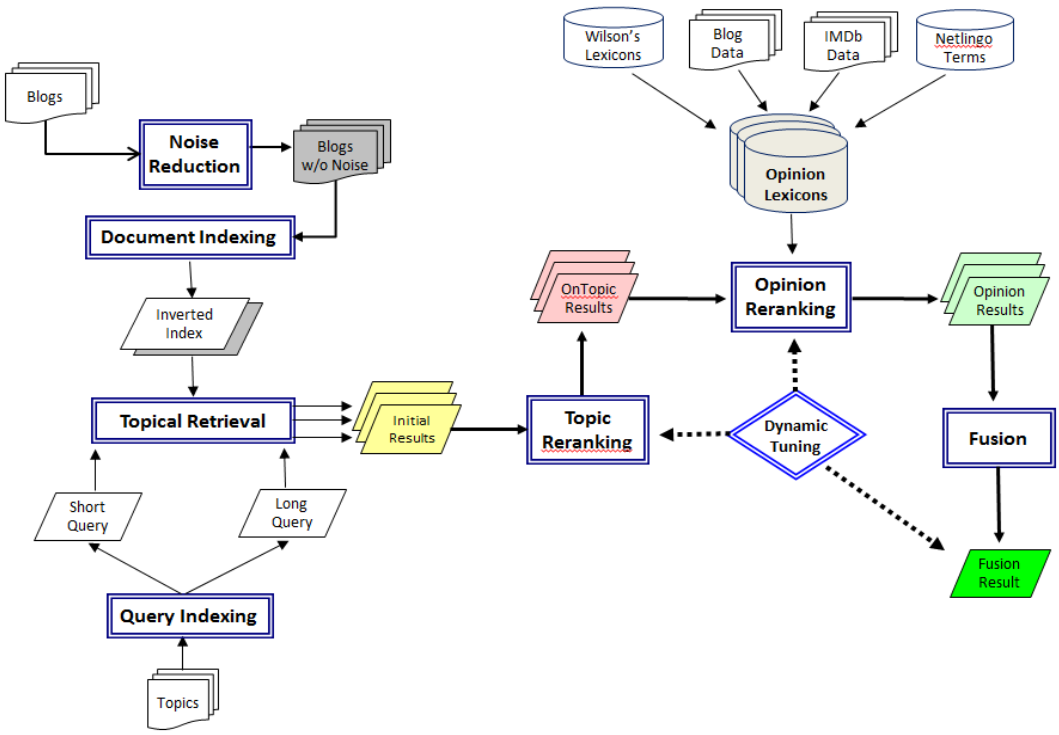
make yet another type of opinion evidence that supplements opinion lexicon and opinion morphology. Opinion collocations give contextual evidences of opinion to otherwise neutral statements that they annotate (e.g., “I think tomato is a fruit”, “Tomato is a vegetable to me”).

As for the association of opinion to its intended target, we hypothesize that the proximity of opinion expression to the target in question can serve as a reasonable surrogate for more complex syntactic analysis of natural language processing techniques. Since we do not apply any anaphor resolution, we believe that proximity scores should be supplemented with non-proximity scores for completeness.

Last but certainly not least, the fusion hypothesis underlies various layers of our approach, from query expansion and topical retrieval to opinion detection and system optimization. The fusion hypothesis, which posits that the whole is greater than sum of its parts, regards the state of “individual weaknesses” and “complementary strengths” as an ideal condition for fusion (Yang 2014). Individual weaknesses describe a state with no single overwhelmingly effective solution to a given problem, while complementary strengths exist if those individual solutions are distinct enough to produce a significantly improved outcome when combined. For instance, three sources of opinion evidence utilized in our opinion detection approach, as well as multiple query expansion methods and variations in opinion scoring formulas are motivated by the fusion hypothesis.

Our approach to targeted opinion detection consists of a two-step process: a topical retrieval process and an opinion detection process. The topical retrieval component reranks an initial topic search for rank optimization to distill documents about a target topic at top ranks, after which the opinion scores are computed by multiple opinion scoring modules that leverage different sources of opinion evidence to maximize the coverage of opinion expressions. In a prior study, we implemented *Opinion Term Module* to make use of the opinion lexicon, *Rare Term Module* to leverage the opinion morphology, *IU Module* to utilize the opinion collocations, and *Adjective-Verb Module* to supplement the three lexicon-based methods with the computational linguistics’ distribution similarity approach (Yang et al. 2007a; 2007b; 2008; Yang 2009). Based on the findings that suggested the advantage of lexicon-based opinion detection methods over classification-based ones, we extended our prior approach by modifying the opinion detection component to be entirely lexicon-based while experimenting with probabilistic weights for lexicon terms and distance-based opinion scores of documents. Figure 1 displays the architecture of our blog opinion retrieval approach, which is implemented in the

WIDIT[1] system.



<Fig. 1> Targeted Opinion Detection System Architecture

1. Noise Reduction

The blogosphere contains three types of noise: spam for the purpose of link farming or advertisement (Adamic and Glance 2005; Efron 2004), non-English blogs, and non-posting contents generated by blogware. In investigating the blog noise question, however, we did not include spam detection so as to isolate the effects of blog-specific noise from the spam effect. In order to assess the effect of blog noise, we devised a noise reduction module that identifies non-English documents and blogware-generated noise for exclusion so that the results of opinion finding runs with and without noise can be compared.

1) WIDIT (Web Information Discovery Integrated Tool, http://widit.knu.ac.kr/) is a research infrastructure constructed and maintained by the author.

1.1 Non-English Blog Identification

To identify non-English (NE) blogs, we first made use of the language tags in blog feeds.[2] Since the feed markup syntax is not standardised, we manually identified NE language tags from all unique tags in the feed data. Blog permalinks listed in the feeds with NE language tags were flagged as NE blogs. To compensate for the feeds without language tags, we formulated a token-based strategy that identifies NE blogs based on the proportion of non-ASCII character tokens (i.e., NE tokens) in a document.

The basic heuristic of the NE token method, which is to flag a blog with a large proportion of NE tokens as a NE document, was refined iteratively by examining the results of the method to eliminate false positives (e.g., English blogs with NE tokens) and reduce false negatives (e.g., NE blogs with English tokens). The final version of the NE token method considers additional factors, such as the frequency and proportion of stopwords and document length, to accommodate NE blogs with English tokens and English blogs with NE tokens.

1.2 Non-Posting Content Exclusion

A blog contains blogware-generated content for navigation, advertisement, and display formatting as well as user-authored content of original posting and follow-up comments, all of which are delineated by markup tags. Unfortunately, non-standard use of different markup tags adopted by different blogwares makes tag-based exclusion of non-posting content (NPC) a rather challenging task. Our approach to NPC exclusion is to extract blog segments based on content-bearing markup tags first to prevent inadvertent exclusion of true content and then to apply noise exclusion to the remaining text based on “noise” tags. Since the tags come in all shapes and sizes and can contain various attributes (e.g., <!—maincontentsarea-->, , <td id=”main_content” width=”466”>, <td class=”sidebar” rowspan=5>,), we constructed regular expressions to identify content and noise tags as described below.

1. Extract all unique tag patterns from blogs.
2. Rank the tags by their occurrence frequency.

---

2) A blog feed contains a summary list of blog entries (i.e., permalinks).

3. Examine the top k tags to compile a list of content and noise tags.
4. Construct regular expressions (regex) to identify the tags in the list.
5. Apply regex to the unique tag set.
6. Refine regex based on the examination of the regex results.
7. Repeat steps 5 & 6 until noise regex produces no false positives (i.e., no real content is excluded) and the majority of high frequency tags are matched by the regex.

2. On-Topic Retrieval Optimization

Optimizing the ranking of the initial retrieval results is important for two reasons. First, on-topic retrieval optimization is an effective strategy for incorporating topical clues not considered in initial retrieval (Yang & Albertson 2004; Yang et al. 2005; 2007a). Second, our two-step strategy for targeted opinion detection consists of minimalistic initial retrieval that favors recall followed by post-retrieval reranking to boost precision.

Our on-topic retrieval optimization involves reranking of the initial retrieval results based on a set of topic-related reranking factors, where the reranking factors consist of topical clues not used in initial ranking of documents. The topic reranking factors used the study are: Exact Match, which is the frequency of exact query string occurrence in document, Proximity Match, which is the frequency of padded[3] query string occurrence in document, Noun Phrase Match, which is the frequency of query noun phrases occurrence in document, and Non-Rel Match,[4] which is the frequency of non-relevant nouns and noun phrase occurrence in documents. All the reranking factors are normalized by document length. The on-topic reranking method consists of following three steps:

1. Compute topic reranking scores for top N results.
2. Partition the top N results into reranking groups based on the original ranking and a combination of the most influential reranking factors. The purpose of reranking groups is to prevent excessive influence of reranking by preserving the effect of key ranking factors.
3. Rerank the initial retrieval results within reranking groups by the combined reranking score.

3) “Padded” query string is a query string with up to *k* number of words in between query words.

4) Non-rel Match is used to suppress the document rankings, while other reranking factors are used to boost the rankings.

The objective of reranking is to float low ranking relevant documents to the top ranks based on post-retrieval analysis of reranking factors. Although reranking does not retrieve any new relevant documents (i.e. no recall improvement), it can produce high precision improvement via post-retrieval compensation (e.g. phrase matching). The key questions for reranking are what reranking factors to consider and how to combine individual reranking factors to optimize the reranking effect. The selection of reranking factors depends largely on the initial retrieval method since reranking is designed to supplement the initial retrieval. The fusion of the reranking factors can be implemented by a weighted sum of reranking scores, where the weights represent the contributions of individual factors. The weighted sum method is discussed in more detail in the fusion section of the methodology.

3. Opinion Detection

Determining whether a document contains an opinion is somewhat different from classifying a document as opinionated. The latter, which usually involves supervised machine learning, depends on the document's overall characteristic (e.g., degree of subjectivity), whereas the former, which often entails the use of opinion lexicons, is based on the detection of opinion evidence. At the sentence or paragraph level, however, the distinction between the two becomes inconsequential since the overall characteristic is strongly influenced by the presence of opinion evidence.

For opinion mining, which involves opinion detection rather than opinion classification, opinion assessment methods are best applied at subdocument (e.g., sentence or paragraph) level. At subdocument level, the challenges of machine learning approach are compounded with two new problems: First, the training data with document-level labels is likely to produce a classifier not well suited for subdocument level classification. Second, the sparsity of features in short "documents" will diminish the classifier's effectiveness. The opinion detection challenge for machine learning was showcased in the TREC blog track, where machine learning approaches to the opinion finding task met with only marginal success (Joshi, Bayrak and Xu 2007; Zhang and Yu 2007; Zhang and Zhang 2007).

Our opinion detection approach, which is entirely lexicon-based to avoid the pitfalls of the machine learning problems, relies on a set of opinion lexicons that leverage various evidences

of opinion. The key idea underlying our method is to combine a set of complementary evidences rather than trying to optimize the utilization of a single source of evidence. In keeping with the main research question of the study, we leveraged the complementary opinion evidences of *Opinion Lexicon* (e.g., suck, cool), *Opinion Collocation* (e.g., I believe, to me), and *Opinion Morphology* (e.g., sooooo, metacool) in the form of semi-automatically constructed opinion lexicons. Opinion lexicons are utilized by opinion scoring modules to compute opinion scores of documents, and the combined opinion score in conjunction with the on-topic score is used to boost the ranks of opinionated documents in a manner similar to the on-topic retrieval optimization.

Opinion scoring modules used in this study are *High Frequency module,* which identifies opinions based on the frequency of opinion terms (i.e., terms that occur frequently in opinionated blogs), *Low Frequency module,* which makes use of uncommon/rare terms (e.g., "sooo good") that express strong sentiments, *IU module,* which leverages n-grams with IU (I and you) anchor terms (e.g., I believe, You will love), *Wilson's lexicon module,* which uses a collection-independent opinion lexicon composed of a subset of Wilson's subjectivity terms, and *Opinion Acronym module,* which utilizes a small set of opinion acronyms (e.g., imho), Each module computes three types of opinion scores for each lexicon used: a simple frequency-based score, a proximity score based on the frequency of lexicon terms that occur near the query string in a document, and a distance-based score computed as a sum of inverse word distances between lexicon terms and the query string. The generalized formula for opinion scoring can be described as

$$opSC(d) = \frac{\sum_{t \in L \cap D} f(t) \cdot s(t)}{len(d)} \quad (2)$$

where L and D denote the term sets of a given lexicon and document d respectively, $len(d)$ is the number of tokens in d, $s(t)$ is the strength of term t as designated in the lexicon, and $f(t)$ is the frequency function that returns either the frequency of t in d (simple score), the frequency of t that co-occurs with the query string in d in a fixed-size window (proximity score), or the sum of inverse word distances between occurrences of t and the nearest query string (distance score). The proximity score, which is a strict measure that ensures the opinion found is on target, is liable to miss opinion expressions located outside the proximity window

as well as those near the target that is expressed differently from the query string. The simple score, therefore, can supplement the proximity score, especially when used in conjunction with the on-topic optimization. A detailed description of the opinion modules can be found in our prior paper (Yang 2009).

4. Fusion

Topic reranking for on-topic retrieval optimization and opinion reranking for opinion detection generate multitudes of reranking scores that need to be combined. Two most common fusion methods are *Similarity Merge* (Fox & Shaw 1995) and *Weighted Sum* (Bartell, Cottrell & Belew 1994). *Similarity Merge,* based on the assumption that documents with higher overlap are more likely to be relevant, multiplies the sum of fusion component scores for a document by the number of fusion components that retrieved the document (i.e. overlap),. Instead of relying on overlap, the *Weighted Sum* method sums the fusion component scores weighted by their relative contributions, which is usually estimated from training data.

In our investigations (Yang 2002; Yang 2014), we found the normalized weighted sum formula to be most effective in combining fusion components that are dissimilar. The normalized weighted sum formula (equation 3) linearly combines the min-max normalized score of component *i*, (nSC_i) with fusion weight w_i, to generate the fusion score. In the min-max normalization (Lee 1997), described in equation 4, $S_i(d)$ is the raw score of document d by component i, and $min\{S_i\}$ and $max\{S_i\}$ are the minimum and maximum scores by the component *i*.

$$FS(d) = \sum_{i=1}^{k} (w_i * NS_i(d)) \tag{3}$$

$$NS_i(d) = \frac{S_i(d) - \min\{S_i\}}{\max\{S_i\} - \min\{S_i\}} \tag{4}$$

In reranking, the original scores should be combined with fusion scores of reranking factors in a way that enables the documents with high reranking factors to float to the top without unduly influencing the existing document ranking. This reranking strategy can be expressed as

$$RS(d) = \alpha * NS_{orig}(d) + \beta * \sum_{i=1}^{k} (w_i * NS_i(d)) \quad (5)$$

where $NS_{orig}(d)$ is the min-max normalized score of document d before reranking, and α and β are the weights that represent the estimated contributions of the original and combined reranking factor scores.

5. Dynamic Tuning

To optimize the reranking formulas, which involves determination of fusion weights (w_i) as well as original and reranking score weights (α and β), we implemented an interactive system tuning interface called *Dynamic Tuning* that displays the effects of tuning parameter changes in real time to guide the human towards the system optimization state in a manner similar to bio-feedback (Yang 2009). *Dynamic Tuning,* which is motivated by the idea of combining human intelligence, especially the pattern recognition ability, with the computational power of the machine, is implemented in a Web application that allows the human to examine not only the immediate effect of his/her system tuning but also the possible explanation of the tuning effect in the form of data patterns. By engaging in iterative dynamic tuning process that successively fine-tune the reranking parameters based on the cognitive analysis of immediate system feedback, system performance can be improved without resorting to an exhaustive evaluation of parameter combinations, which can not only be prohibitively resource intensive with numerous parameters but also fail to produce the optimal outcome due to its linear approach to factor combination.

Ⅴ. Experiment

We tested our targeted opinion detection approach with TREC's Blog06 corpus, 50 topics (Topic 901-950) that describe the information needs for the opinion finding task, and associated relevance judgments. By first applying topic reranking to initial retrieval results and then applying opinion reranking, we generated a result set of 1000 blogs for each topic, which

were evaluated and analysed to assess the effectiveness of our methodology.

1. Data

The Blog06 corpus includes a crawl of feeds (XML), associated permalinks (blog postings in HTML), and feed homepages captured from December 2005 through February 2006. Among the blog document set of 100,649 feeds (39GB), 3.2 million permalinks (89GB), and 325,000 homepages (21GB), only the permalinks were judged for relevance in the opinion finding task. TREC assessed a pool of unique results created from merging top ranked blogs from submitted results. To be considered relevant, a blog had to be on topic and contain an explicit expression of opinion or sentiment about the topic, showing some personal attitude either for or against. Each of 50 test topics, which were constructed by TREC using queries from commercial blog search engines (e.g., BlogPulse and Technorati), consists of title (phrase), description (sentence), and narrative (paragraph) fields. For our study, we generated short queries from the title field and long queries from all three fields. An example of the TREC blog topic is shown in figure 2.

<title> "howard stern" </title>

<desc> Description:
Find opinions of shock jock Howard Stern.
</desc>

<narr> Narrative:
Opinions of Howard Stern and his radio program are relevant. Comments and opinions expressed by Stern are not relevant. Opinions expressed by guests on the program are relevant only if about Stern or the program but not if about the program's topic.
</narr>

〈Fig. 2〉 Topic 922 of TREC Blog track Opinion Finding task

2. Evaluation Metrics

The main performance evaluation metric used in the study is mean average precision (MAP), which is the average precision averaged over all topics. The average precision, which is computed by averaging the precision values obtained after each relevant document is retrieved, is a single-valued measure that considers both recall and precision to gauge the

overall performance. Mean R-precision (MRP), which is the precision at rank R (total number of relevant items) averaged over topics, and precision at rank N (P@N) were also used to evaluate the system performances. Mean R-accuracy, which is the fraction of retrieved documents above rank R that are correctly classified averaged over topics, is used for evaluating polarity subtask results. R-precision and R-accuracy are measures that aim to produce more robust cross-system evaluations by dampening the effects of exact document rankings and variance of R across topics.

Ⅵ. Results

As described in the methodology section, our approach to targeted opinion detection combined topic reranking to optimize the on-topic retrieval, opinion reranking to integrate opinion detection into on-topic retrieval, and fusion at various levels to affect the complementary combination of multiple methods as well as sources of evidence.

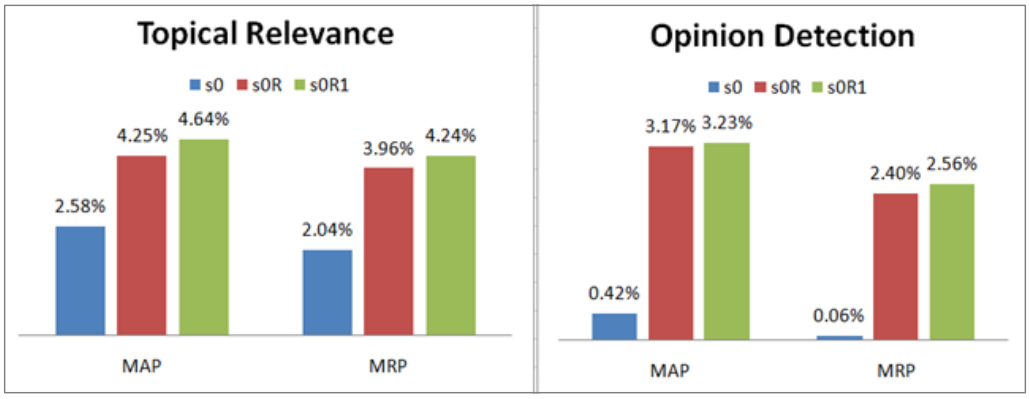
1. Noise Reduction

The feed language tag method of our NE identification module flagged 16,121 blog permalinks while the NE token method flagged 334,219 permalinks as non-English documents. The result of NE identification was validated by an exhaustive review of 2,304 documents flagged as NE in the 2006 relevance judgments, which found only 3 false positives of little consequence (blogs with empty or gibberish content). The review also revealed 21 documents that were judged as topically relevant and 3 that were judged as opinionated, all of which appeared to be non-English documents that constitute potential errors in the relevance judgment file.

While NE identification excluded 12% of the blog permalinks, non-posting content (NPC) exclusion reduced noise levels in over 50% of the permalinks, where the average length reduction was 551 bytes. Overall, 74.3 million (7.4 %) tokens were excluded by NPC method. In terms of unique tokens, the NPC exclusion amounted to 21,283 (0.6 %) terms, which suggests the existence of common noise patterns. Table 1 compares the blog permalink statistics before and after NPC exclusion.

<Tab. 1> Blog Permalink Statistics before and after Noise Reduction

	With Noise	Without Noise
Token count	1,001,269,418	926,967,569
Unique token count	3,343,474	3,322,191
Avg. blog length (byte)	4,299	4,019
Avg. blog length (unique terms)	227	210



<Fig. 3> Noise Reduction Effects

In order to isolate the effect of noise reduction on system performance, we examined the performance differences between retrieval run pairs that are identical in all aspects except for noise reduction. Figure 3 displays average pairwise differences in performance by noise reduction in various categories. The upward direction of the bars indicates the gain in performance by noise reduction, while blue bars represent the initial retrieval results without reranking (s0) and green and red bars indicate reranked results with (s0R1) and without dynamic tuning (s0R). The beneficial effect of noise reduction is clearly illustrated in Figure 4, where the performance gain by noise reduction is consistently positive across evaluation metrics as well as types of evaluations (i.e., topical relevance, opinion detection, polarity classification). Subsequent discussions of the result are restricted to those with noise reduction applied.

2. Reranking

Table 2 and 3 show the topic and opinion MAPs of the reranked results in comparison with

the initial retrieval results (i.e. baseline). Topic MAP is the MAP computed using only the topical relevance (i.e., document is topically relevant), whereas opinion MAP is computed using the opinion relevance (i.e., document is topically relevant and opinionated).

〈Tab. 2〉 Topic MAP

	Short Query	Long Query	Fusion
Baseline	.3715	.3736	.4065 (+8.8%)
Topic Rerank	.4088 (+10.0%)	.4062 (+8.7%)	.4290 (+5.6%)

〈Tab. 3〉 Opinion MAP

	Short Query	Long Query	Fusion
Baseline	.2772	.2817	.3027 (+7.5%)
Topic Rerank	.2902 (+4.7%)	.2984 (+5.9%)	.3086 (+3.4%)
Opinion Rerank	.3272 (+12.7%)	.3369 (+13.0%)	.3408 (+1.2%)

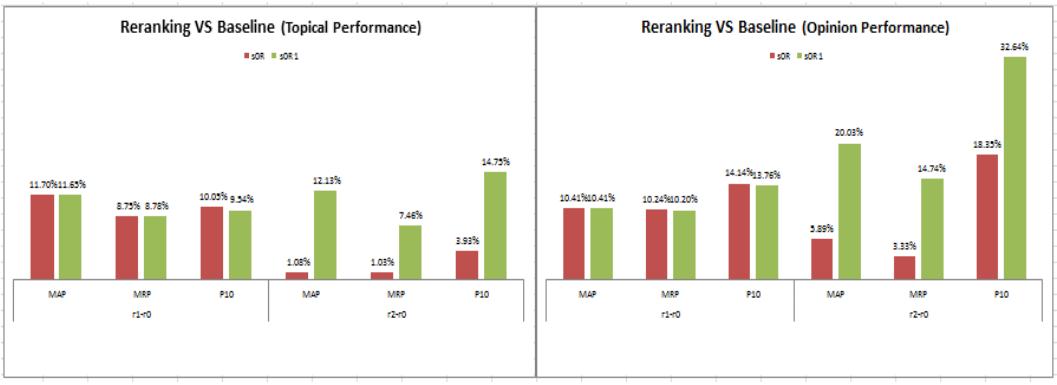
Numbers in Tables 2 and 3 indicate that both topic and opinion reranking strategies are working as intended. Topic reranking effectively improves the on-topic retrieval performance of the baseline retrieval (10.0% for short and 8.7% for long query), after which opinion reranking boosts the o1pinion finding performance even more strongly (12.7% and 13.0%). It is interesting to note that the op_MAP improvements by topic rereanking are only about half of tp_MAP improvements (4.7% vs. 10.0% for short and 5.9% vs. 8.7% for long query) and tp_MAP improvements by opinion reranking are less than half of op_improvements (5.6% vs. 12.7%, 6.0% vs. 13.0%). This suggests that the relationship between topical and opinion relevance is not strictly linear. Topical relevance and opinion relevance may be correlated but the degree of topical relevance and the degree of opinion relevance may not be. In other words, a topically relevant document is more likely to contain opinions about the topic than topically irrelevant documents but a high ranking topically relevant document is not necessarily more likely to be opinionated than low ranking topically relevant documents.

Fusion of reranked documents, on the other hand, affects only marginal performance improvements. The improvements by fusion over the best non-fusion results are 5.6% tp_MAP for topic reranking and 1.2% op_MAP for opinion reranking. Possible explanations for the muted effect of fusion are: one, we did not engage in dynamic tuning to optimize the fusion

formula as were done with reranking formulas; two, reranking produced near-optimal results with implicit fusion of reranking factors, thus leaving less room for improvement by explicit fusion of reranked results.

3. Dynamic Tuning

Figure 4, which displays average performance improvements over baseline by topical and opinion relevance, shows the effect of topic reranking on the left side (r1-r0) and opinion reranking effect on the right side (r2-r0). The green bars (s0R1) indicate dynamic tuning results, while the red bars (s0R) represent results without dynamic tuning. The average performance improvement over baseline is computed by averaging the performance differences between all system pairs that are identical in all aspects except for the parameter in question (e.g., baseline vs. topic reranking with dynamic tuning). The fact that all the bars point upward shows that reranking is generally beneficial by all three measures (MAP, MRP, P@10). The markedly taller green bars on the right side suggests that dynamic tuning is quite effective in optimizing the opinion reranking formula.



〈Fig. 4〉 Dynamic Tuning Effect

4. Comparative Performance

Table 4 shows the performance of our current system (WIDIT2[5]) relative to TREC

5) WIDIT in Tables 4 and 5 refers to our old system used in TREC-2007 participation.

participant performances. What should be noted at this point is how much of the final performance is due to initial retrieval and how much of it is by the opinion detection strategy. Our baseline performance is far below the top TREC system. In fact, op_MAP of *UAmsterdam* is that of the baseline run. To assess the effectiveness of opinion detection methods, we need to compare baseline and opinion detection results of systems side by side (Table 5).

Opinion detection performances of all TREC-2007 participants are shown in Figure 5. Each line, representing a TREC participant, starts on the left with the MAP of his/her baseline system and ends on the right with the MAP of his/her opinion detection system. The group of lines on the left shows participants whose opinion detection strategy had rather severely detrimental effects on performance, the middle lines show those with negligible opinion detection effects, and the lines on the right side show the ones with effective opinion detection methods. The fact that only one third of the TREC participants were able to devise effective opinion detection strategies reflects the difficulty associated with the targeted opinion detection task.

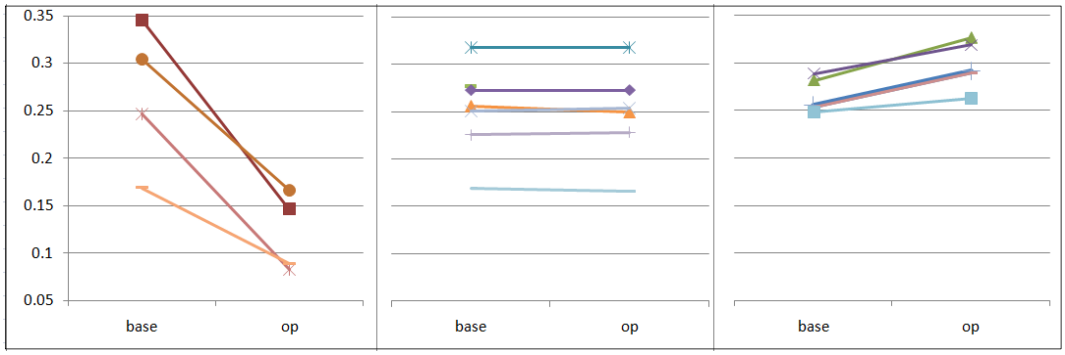
<Tab. 4> Relative Short Query Performance of WIDIT Opinion Finding System

System	Rank	Best MAP	Baseline	Baseline Rank
UIUC	1	.4341	?[6]	?
UAmsterdam	2	.3453	.3453	1
WIDIT2	**3**	**.3271**	**.2772**	**6**
WIDIT	8	.2894	.2537	10

<Tab. 5> Improvements over Baseline for Short Query

System	Rank	Baseline MAP	Opinion MAP	% Increase
WIDIT2	1	.2772	.3271	18.0%
UGlasgow	2	.2817	.3264	15.9%
WIDIT	3	.2537	.2894	14.1%

6) The baseline performance of UIUC is unknown since they did not submit a baseline run.



<Fig. 5> Opinion Detection Performances of TREC participants

Ⅶ. Concluding Remarks

In this paper, we presented a fusion approach to finding opinionate blogs on a given target. Our approach to finding opinionated blogs on target consists of first generating a good topical search result and then boosting the ranks of opinionated documents. On-topic retrieval optimization considers topical factors not utilized in the initial search to increase precision in the topical retrieval result, after which the opinion evidences in documents are leveraged to optimize the ranks of opinionated documents. Both on-topic retrieval optimization and opinion detection involve reranking of documents based on the combination of multiple reranking factors. In combining reranking factors, we used a weighted sum formula in conjunction with a reranking heuristic that are iteratively tuned via an interactive system optimization process called Dynamic Tuning.

Our opinion finding system in TREC-2006 blog track produced the best result among participants, but our results in TREC-2007 dropped to the 8th rank for the short query. We attribute our mediocre performance in the short query category in 2007 to the poor baseline result, which was at rank 10 among participants. In this study, we refined our noise reduction module to short queries to improve the baseline performance and extended our 2007 opinion module by combining the manual lexicon weights with probabilistic weights, and experimenting with distance-based scoring formula, all of which increased our short query performance by 13% (from .2894 to 0.3271) to move up to the rank 3.

Our experimental results, which were produced by applying our targeted opinion detection system in a standardized environment of TREC, clearly demonstrated the effectiveness of

combining multiple complementary lexicon-based methods for opinion detection. The analysis of the results also revealed that Dynamic Tuning is a useful mechanism for fusion, and post-retrieval reranking is an effective way to integrate topical retrieval and opinion detection as well as to optimize the baseline result by considering factors not used in the initial retrieval stage. In comparison to top TREC blog track systems, our targeted opinion detection approach performed not only competitive but also quite effective in leveraging opinion evidences to improve the baseline performance.

References

Lada, Adamic and N. Glance. 2005. "The political blogosphere and the 2004 US election: Divided they blog." *Proceedings of the 3rd International Workshop on Link discovery,* 36-43.

Bartell, Brian T., G. W. Cottrell and R. K. Belew. 1994. "Automatic combination of multiple ranked retrieval systems." *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval,* 173-181.

Chklovski, Timothy. 2006. "Deriving quantitative overviews of free text assessments on the web." *Proceedings of the 11th International Conference on Intelligent User Interfaces,* 155–162.

Ding, Xiaowen., B. Liu and P. S. Yu. 2008. A holistic lexicon-based approach to opinion mining. *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 231-240.

Efron, Miles. 2004. "The liberal media and right-wing conspiracies: using cocitation information to estimate political orientation in web documents." *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, 390–398.

Fox, Edward A. and J. A. Shaw. 1995. "Combination of multiple searches." *Proceeding of the 3rd Text Retrieval Conference*, 105-108.

Hu, Minqing and B. Liu. 2004. "Mining and Summarizing Customer Reviews." In *KDD'04: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177.

Joshi, Hemant, C. Bayrak and X. Xu. 2007. "UALR at TREC: Blog Track." *Proceedings of the 15th Text Retrieval Conference*.

Lee, Joon H. 1997. "Analyses of multiple evidence combination." *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 267-276.

Liu, Bing, M. Hu and J. Cheng. 2005. "Opinion observer: analyzing and comparing opinions on the web." *Proceedings of the 14th International Conference on World Wide Web*, 342–351.

Liu, Bing. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 1-167.

Macdonald, Craig, R. L. Santos, I. Ounis and I. Soboroff. 2010. Blog track research at TREC. *ACM SIGIR Forum*, 44(1), 58-75.

Mishne, Gilad. 2007. "Multiple Ranking Strategies for Opinion Retrieval in Blogs." *Proceedings of the 15th Text Retrieval Conference*.

Mishne, Gilad and M. de Rijke. 2006. "Deriving wishlists from blogs: Show us your blog, and we'll tell you what books to buy." *Proceedings of the 15th International World Wide Web Conference*. 925-926.

Oard, Doug, T. Elsayed, J. Wang, Y. Wu, P. Zhang, E. Abels and D. Lin. 2007. "TREC 2006 at Maryland: Blog, Enterprise, Legal and QA Tracks." *Proceedings of the 15th Text Retrieval Conference*.

Ounis, Iadh, C. Macdonald, J. Lin and I. Soboroff. 2011. Overview of the TREC-2011 microblog track. *Proceedings of the 20th Text Retrieval Conference (TREC 2011)*.

Ounis, Iadh, C. Macdonald, M. de Rijke and G. Mishne. 2007. "Overview of the TREC 2006 Blog Track." *Proceedings of the 15th Text Retrieval Conference*.

Taboada, Maite, J. Brooke, M. Tofiloski, K. Voll and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267-307.

Thelwall, Mike, K. Buckley and G. Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1), 163-173.

Thelwall, Mike, K. Buckley, G. Paltoglou, D. Cai and A. Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society*

for Information Science and Technology, 61(12), 2544-2558.

Wiebe, Janyce, T. Wilson, R. Bruce, M. Bell and M. Martin. 2004. "Learning subjective language." *Computational Linguistics*, 30(3), 277–308.

Wilson, Theresa, D. R. Pierce and J. Wiebe. 2003. "Identifying opinionated sentences." *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 33–34.

Yang, Kiduk. 2002. "Combining Text- and Link-based Retrieval Methods for Web IR." *Proceedings of the 10th Text Retrieval Conference*, 609-618.

Yang, Kiduk. 2009. WIDIT in TREC 2008 Blog Track: Leveraging Multiple Sources of Opinion Evidence. *Proceedings of the 17th Text Retrieval Conference*.

Yang, Kiduk. 2014. Combining multiple sources of evidence to enhance Web search performance. *Journal of Korean Library and Information Science Society*, 45(3), 5-36.

Yang, Kiduk and N. Yu. 2005. "WIDIT: Fusion-based approach to Web search optimization." *Information Retrieval Technology*, 206-220. New York: Springer-Verlag.

Yang, Kiduk, N. Yu, A. Valerio and H. Zhang. 2007a. "WIDIT in TREC2006 Blog track." *Proceedings of the 15th Text Retrieval Conference*.

Yang, Kiduk, N. Yu, A. Valerio, H. Zhang and W. Ke. 2007b. "Fusion approach to finding opinions in Blogsophere." *Proceedings of the 1st International Conference on Weblog and Social Media*.

Yang, Kiduk, N. Yu, A. Wead, G. La Rowe, Y. H. Li, C. French and Y. Lee. 2005. "WIDIT in TREC2004 Genomics, HARD, Robust, and Web tracks." *Proceedings of the 13th Text Retrieval Conference*.

Yang, Kiduk, N.Yu and H. Zhang. 2008. "WIDIT in TREC2007 Blog track: Combining lexicon-based methods to detect opinionated Blogs." *Proceedings of the 16th Text Retrieval Conference*.

Zhang, Ethan and Y. Zhang. 2007. "UCSC on TREC 2006 Blog Opinion Mining." *Proceedings of the 15th Text Retrieval Conference*.

Zhang, Wei and C. Yu. 2007. "UIC in TREC 2006 Blog Track." *Proceedings of the 15th Text Retrieval Conference*.