A Combined Query Expansion Technique for Retrieving Opinions from Blogs

Saeedeh Momtazi, Stefan Kazalski, Dietrich Klakow
Spoken Language Systems
University of Saarland
Saarbruecken, Germany
{saeedeh.momtazi-stefan.kazalski-dietrich.klakow}@lsv.uni-saarland.de

Abstract

In this paper, we discuss the the role of the retrieval component in an TREC style opinion question answering system. Since blog retrieval differs from traditional ad-hoc document retrieval, we need to work on dedicated retrieval methods. In particular we focus on a new query expansion technique to retrieve people's opinions from blog posts. We propose a combined approach for expanding queries while considering two aspects: finding more relevant data, and finding more opinionative data. We introduce a method to select opinion bearing terms for query expansion based on a chi-squared test and use this new query expansion to combine it in a liner weighting scheme with the original query terms and relevant feedback terms from web. We report our experiments on the TREC 2006 and TREC 2007 queries from the blog retrieval track. The results show that the methods investigated here enhanced mean average precision of document retrieval from 17.91% to 25.20% on TREC 2006 and from 22.28% to 32.61% on TREC 2007 queries.

1 Introduction

The processing of opinion information has been widely discussed nowadays, because humans like to express their opinions on the Internet and are eager to let others know about their opinions. Motivation for this task comes from the desire to provide tools to analyze this information for individuals, governmental organizations, commercial companies, and political groups, who want to automatically track attitudes and feelings in on-line resources. What do students like about Wikipedia? How do people feel about recent events in the Middle East? Who likes Microsoft products? What organizations are against universal health care? What are the public opinions on human cloning? What users prefer Google Mail?

Blog data is one of the most prevalent sources among others that provided opinionated documents. The rise on blogs has provided a new subset of the World Wide Web that represents real-world events. Since the number of blog writers and readers rapidly increases, blog pages become an increasingly important information source about people's personal ideas, beliefs, feelings, and sentiments (positive or negative). Indeed, such subjective information in blog pages is useful to find out what people think about various topics in making decisions. Hence it opens up several new interesting research areas.

A system that could automatically identify opinions and emotions from text would be an enormous help to someone trying to answer these kinds of questions. Natural language processing applications could benefit from being able to distinguish between factual and opinionative information. Question answering systems which can detect and classify factual and opinionative information offers distinct advantages in deciding what information to extract and how to organize and present this information. Such system aims to present multiple answers to the user based upon opinions derived from blogs. Most of the state-of-the-art question answering systems focus factual questions. However, opinion questions have longer and more complex answers. The answers tend to be scattered across different documents. Traditional question answering approaches are not effective enough to retrieve answers for opinion questions as they have been designed for factual questions. Hence, an opinion question answering system is different from a factoid question answering system and in particular needs dedicated retrieval algorithms.

In this research, we improve our current question answering system to deal with opinion questions and extract their answers from blogs. Since the document retrieval component is the major part of a question answering system which should be able to retrieve opinions from blogs, we focus on this component to find the best answers for this

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.196

kind of questions. A good retrieval system means that only small number of top ranked documents needs to be analyzed by the answer extraction in order to find the answer. In this paper, we improve blog document retrieval within the question answering context by proposing a new query expansion technique.

The rest of this paper is structured as follows: In the next section, we describe the previous work on opinion retrieval. Section 3 talks about our document retrieval module. In Section 4, we introduce our proposed method for query expansion, and Section 5 presents our results. Finally, Section 6 concludes the paper with a summary.

2 Related Work

There has been a spate of research on identifying opinion in document and sentence retrieval and especially in QA systems.

Wiebe [14] proposed a method to identify strong clues of subjectivity on adjectives. He introduced subjectivity tagging for distinguishing sentences used to present opinions from sentences used to present factual information. At document level, Wibe [16] recognized opinionative documents by demonstrating a straightforward method for learning certain kinds of potentially subjective collocations from corpora. A year after, Weibe [15] continued his research by proposing a method for opinion summarization. Riloff and Wiebe [12] presented a subjectivity classifier using lists of subjective nouns learned by bootstrapping algorithms to distinguish opinionative sentences from factual ones. In the first step, they used two bootstrapping algorithms that exploit patterns extraction to learn sets of opinion nouns. Then, they trained a Naive Bayes classifier using the subjective nouns, discourse features, and subjectivity clues. In another research, they [11] proposed a bootstrapping process to learn linguistically pattern extraction for subjective expressions. The learned patterns are then used to identify more subjective sentences. In 2005 they expanded their research by working on Multi-Perspective Question Answering (MPQA) systems. As an initial step towards the development of MPQA systems, they investigated the use of machine learning and rule-based subjectivity and opinion source filters and showed that they can be used to guide MPQA systems [13].

Pang, Lee, and Vaithyanathan [10] classified documents by overall sentiments instead of topics, and then determined the polarity of a review. Pang [9] proposed a novel machine learning method that applies text categorization techniques to the subjective portions of the document. In this method, he used some efficient techniques for finding minimum cuts in graphs.

In another research by Mukras [8], different machine learning techniques applied to sentiment classification were compared. He used an original corpus and 5 variations of tagging to train and test three classifiers: the Naive Bayes classifier, the Neural Networks classifier, and the Support Vector Machines (SVM) classifier. He showed that on average, SVM yield the best results when test documents are represented as feature presence vectors and Naive Bayes yields the best average result when test documents are represented as feature count vectors. He also noted that representing test documents as feature presence vectors is more useful in the task of sentiment classification .

Yu and Hatzivassiloglou [17] separated opinions from facts, at both the document and sentence levels. They intended to cluster opinion sentences from the same perspective together and summarize them as answers to opinion questions.

Kim and Hovy [5] presented a sentiment classifier for English words and sentences, which utilizes thesauri to determine word sentiments and combined sentiments within a sentence. However, template-based approach needs a professionally annotated corpus for learning, and words in thesauri are not always consistent in sentiment. Kim and Hovy [6] also identified opinion holders, which are frequently asked in opinion questions.

The blog retrival group of The University of Illinois at Chicago [18] developed a two-step approach that finds relevant blog documents containing opinioned content for a given query topic. The first step, retrieval step, is to find documents relevant to the query. The second step, opinion identification step, is to find the documents containing opinions within the scope of the document set from the retrieval step. In the retrieval step, they improved the retrieval effectiveness by retrieving based on concepts, and doing query expansion using web feedback. In the opinion identification step, they trained a sentence classifier using subjective and objective sentences, which extracted from rateitall.com and wikipedia.com, respectively. A year after, this group [19] expanded their system adding a new step, ranking step, which identifies the query-related opinions in the documents and ranks them by computing their opinion similarity scores. They also used a "split and merge" strategy in the polarity task. This strategy is used for finding the positive and negative documents and then find the mixed opinionative documents in the intersection of the positive and negative document sets.

Ernsting et al. [4] used a mixture model of external expansion and document priors to improve opinion finding. They compared their mixture model with Indri in performance. Their best result was achieved by rewriting queries first and then expanding them using an external news corpus. They believe that opinion finding is highly dependent on topical retrieval and opinion detection can be done using lexicons. Their reseach shows that query expansion is one of the promising approaches in opinion finding.

3 Document Retrieval Engine

The task of document retrieval is an important part of a question answering system as it provides the input to the sentence retrieval and the answer extraction. In a question answering system, the document retrieval is used to decrease the number of documents in a large corpus. This is done to reduce the search space in which a correct answer can be found. It is necessary to reduce the search space because the following components (i.e. answer extraction) may use time consuming deep analysis algorithms which strongly depend on the size of the processed corpus. Therefore, it is important to process just the documents which seem relevant to a query and contain opinions about the question to get answers within an appropriate period of time.

The current document retrieval of our question answering system has been built to retrieve factoid data relevant to factoid questions. The extraction of opinionative documents from blogs is not a task that our system as described on can perform without a significant reduction in the quality compared to the relevant documents.

In order to facilitate an improved retrieval and to enable us to implement a query expansion with opionon bearing terms, we switched to INDRI [1] as a retrieval engine which is part of the Lemur Toolkit. It improves language model based retrieval by including inference networks which allows us to flexibly combine the phrases of the query and the expansion terms with weights and connectors while in Lemur all query terms should be considered in the same way. This makes the retrieval more robust against noisy expansion terms.

4 Query Expansion

One of the appropriate methods for retrieving opinions from blogs is using query expansion techniques. By expanding the user's entered terms, more documents are retrieved which increases the system recall at the expense of precision. So, the main goal of query expansion is to increase recall; while precision can potentially increase. Achieving this goal is possible if the expanded query can retrieve more relevant or at least equally relevant documents compared to the original query.

When comparing factoid retrieval to the retrieval of blog documents, opinions enter as a completely new dimension to the problem. Also the document representation and the noise in the collection are different. As a result, we need to find more sophisticated methods for query expansion in blog retrieval than in traditional ad-hoc retrieval which has been studied for a long time. In typical retrieval systems, query expansion is applied to overcome the exact matching problem of the document retrieval and solve the vocabulary mismatches between the query and the document collection. So, the synonyms of the query terms are useful to find relevant documents as they contain different surface realizations of the same entity. In blog retrieval, however, we need to bridge the gap between the information that the users need and the documents likely to be written in blogosphere.

In our task, we expand the query with a set of feedback terms using the web. Expanding the query by feedback terms from the web, however, is no guarantee to retrieve the opinionative documents in high ranks. Using a set of opinion bearing words gives us the chance to retrieve more opinionative documents than before. So, we expand the query with a set of opinion markers. Different researchers used query expansion techniques for retrieving blogs. However, to the best knowledge of the authors, there is no research that simultaneously focused on both approaches for query expansion.

The final set of query terms would be the union of query seed terms, the top k feedback terms and a set of l opinion markers. These set of terms are combined in a weighting schema as follows:

$$\begin{aligned} Q = {} & W_1 \times (\mathrm{QuerySeedTerms}) \\ & + W_2 \times (\mathrm{FeedbackTerms}) \\ & + W_3 \times (\mathrm{OpinionBearingTerms}) \end{aligned} \tag{1}$$

and can be used as input of our document retrieval component.

The scenario of extracting the query seed terms and adding two different groups of terms (feedback terms and opinion bearing terms) for constructing the final query is as follows:

1. **Extracting Query Seed Terms:**
   Noun and verb phrases contained in the query are identified by using Brill's Part of Speech Tagger [3] and Abney's Chunk Parser [1]. The target and the phrases extracted from the query are considered as the set of query seed terms.

2. **Retrieving Feedback Terms:**
   Relevant terms are extracted in four steps. First, each query seed term is sent to web to create a set of feedback documents. Arguello et al. [2] used the *Wikipedia* corpus for their task. However, we found that using both *Wikipedia* and web search engines enhance the performance. So, we send the query seed terms to *Wikipedia* and different web search engines including *Google*, *MSN*, and *Yahoo*. The set of feedback documents are tagged and parsed by the Brill Tagger and the

[1] http://www.lemurproject.org/indri

Charniak Parser in the second step. Then, the degree of relevance of each term of documents is computed as follows:

$$R(t|q_0..q_k) = \sum_{D \in \mathcal{D}} P(D)P(t|D)\prod_{i=0}^{k} P(q_i|D) \quad (2)$$

The relevant term t can be a word or a chunk in our implementation. The terms' relevance score in a document $D \in \mathcal{D}$ is calculated by the probability of the document D and the probability that document contains the term t and the query term q_i while assuming the query terms independent. The total relevance score is the sum of the terms score over all documents. Finally, the terms are ranked in the descending order and the top k terms are selected.

3. **Selecting Opinion Bearing Terms:**
The most informative opinion bearing words for our query expansion task are selected in five steps while using two well-known web pages *http://www.rateitall.com* and *http://www.wikipedia.com* as our resource. In our experiments, *rateitall* is considered as a set of opinionative data and *wikipedia* is used as a set of factoid data. In the first step, we remove the stop words. In the next step, the ch-squared function is used to find the opinionative words. To this end, all documents in *rateitall* are annotated as opinionative data and all documents in *wikipedia* are annotated as factoid data. Having a large number of documents in both classes of opinionative and factoid, we can use the chi-squared function and find the dependecy degree of all terms to each class. Since in this task we need to find the terms that are dependent to the opinionative class, we only calculate the chi-squared of all terms and the opinionative class. Chi-squared of each term in the vocabulary is calculated by the following formula:

$$Chi - Squared(t_i, c) = \frac{N \times (AD - BC)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (3)$$

where t_i is the i^{th} term, c in our case is the opinionative class, and N is the number of documents in the collection. A means the number of times where t_i occur in class c, B means the number of times where t_i occurs without c, C means the number of times where c occurs without t_i, and D means the number of times where neither c nor t_i occur. Having the chi-squared value of all of the terms, in the third step, we rank the words in descending order. The top words mean that they are mostly relevant to the opinionative class and are most likely to occur in an opinionative context. The only problem of this set of words is that there are a lot of proper nouns, specially the name of famous people, which occur frequently with a high rank in the set. The most important reason is having many documents in the opinionative data which present people's opinions about particular famous characters. The names of these characters are usually in the top ranks of the set. As a result, all name entities occurred in the set are removed in the fourth step. Finally, the top l terms are selected to expand the query.

5 Results

The blog data used in this research is the *Blog06* corpus created by the University of Glasgow[7]. The corpus is a collection of homepages and permalinks from blog homepages monitored over an 11 week period from December 2005 to February 2006. The collection contains 3,215,171 homepages. Since blog documents are user created, the texts are short, contain noisy documents with many spelling errors or missing punctuation marks and use uncommon language.

To evaluate our models, we used the set of TREC[2] 2006 and TREC 2007 queries from the Blog Retrieval Track. Each of these query sets contains 50 queries and each query comes with a topic which is often the target word of the query. The relevant blog documents for each query are released by NIST[3], so that for each pair of query topic and blog post the content of the blog post is judged and assigned a label. Table 1 shows the meaning of different labels used by NIST for annotating documents.

Table 1. The Scale Used for Document Annotation

Label	Relevance Scale
0	Not Relevant
1	Relevant - Not Opinionative
2	Relevant - Negative Opinionative
3	Relevant - Mixed Opinionative
4	Relevant - Positive Opinionative

In the first step of query expansion, the seed terms of each query were extracted. Then, using the seed terms the feedback terms were selected. Finally, the top three opinion bearing terms were added to the query. As mentioned before, to put the final query terms in a weighting schema like Equation 1, we need three different values as W_1, W_2, and W_3. In all of our experiments, the following values are used: $W_1 = 2$, $W_2 = 1.5$, and $W_3 = 1$. Inasmuch as

[2] http://trec.nist.gov

[3] National Institute of Science and Technology

the optimum number of feedback terms to be added to the query is not self-evident, tests were conducted with several numbers of feedback terms on *Blog06*. Figure 1 shows the Mean Average Precision (MAP) of the document retrieval for varying numbers of feedback terms. According to the results, the best MAP is achieved by adding 10 feedback terms to each query. Hence, this value was used in all further experiments.

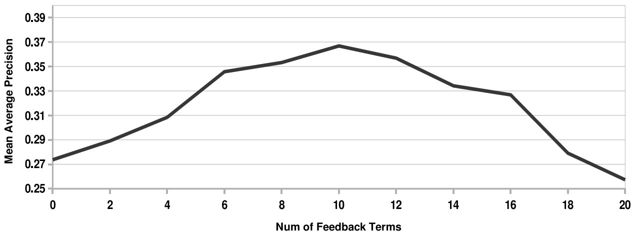

Figure 1. MAP over different numbers of feedback terms

To evaluate our proposed techniques, we used two criteria. In the first one, the retrieved documents were evaluated based on their relevance to the query without considering their opinionativeness. On the other words, all of the documents labeled as 1, 2, 3, or 4 were accepted. Table 2 reports the results of our experiments based on this criterion, in which we use the classic document retrieval with no query expansion as our baseline. The results of both document retrieval (without query expansion and with combined query expansion) are presented on MAP and Precision at level 10 (P@10) as two important factors in information retrieval evaluation.

As presented in the table, on both data sets (TREC 2006 and TREC 2007) the document retrieval with proposed query expansion significantly outperforms the baseline. The results show that the improvement achieved on MAP is more than on P@10, which indicate that our proposed method mostly enhanced the system recall than the system precision. However, all differences are statistically significant at the level of *p-value*< 0.01 based on two tailed *t*-tests.

Table 2. Performance of Document Retrieval on TREC Blog Queries

Queries	No Expansion		Combined Expansion	
	MAP	P@10	MAP	P@10
TREC 2006	0.2737	0.720	0.3668	0.7640
TREC 2007	0.2984	0.7000	0.4292	0.7620

After evaluating the results based on their relevance to the query, we also evaluate them based on the second criterion. In this scenario, the opinionativeness of the documents are important as well as their relevance. So, we only accept the documents that are relevant to the query and also are opinionative. As our question answering system is an opinionative question answering system, it is very essential to retrieve the relevant documents that are opinionative. As a result, our document retrieval component should be evaluated based on the second criterion which considers both aspects. In order to consider the second criterion, we only accepted the documents labeled as 2, 3, or 4; and did not consider the documents labeled as 1.

Figure 2 shows the results of document retrieval based on different forms of query expansion on two different years of TREC data, in which they are evaluated based on both the opinionativeness of documents and their relevance to the query. The first columns present the result of the retrieval system with no query expansion. The second columns show the results of the document retrieval while only the set of feedback terms is used for query expansion. The third columns show the results of the document retrieval in which we only used the set of opinion bearing terms to expand the query. The last columns are the result of our system using both feedback terms and opinion bearing terms for query expansion. The differences between the first and the second columns represent how feedback terms can improve the system performance. The differences between the first and the third columns also indicate that the opinion bearing words alone also have positive effects on the results. The above differences show that each of these two sets plays an important role to improve the performance of our system. However, as we can see from the last columns, we achieved a very significant improvement in the mean average precision by combining both types of query expansion. As it is clear in this figure, although most of the researches only focus on the relevant terms to expand the query in blog retrieval, it is not enough for retrieving opinions in high rank.

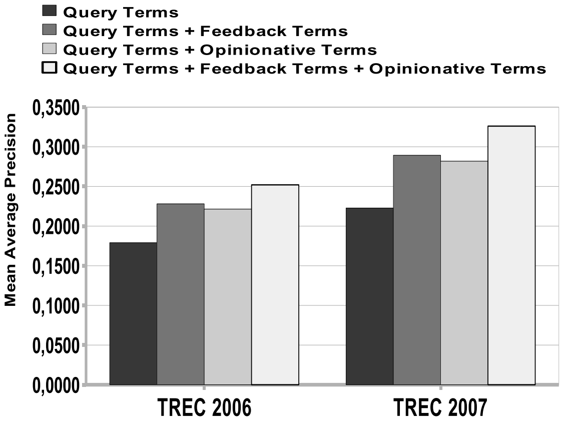

Figure 2. MAP over different query forms

In the development of our question answering system we observed that not only the standard query expansion but also the new combined version has a significant impact on the overall performance of the question answering system as the system has access to more informative data and consequently can extract more accurate answers to the questions.

6 Concluding Remarks

In this paper, we studied query expansion for document retrieval as it can be found as a part of an opinion question answering system. We benefited from blog data to access and retrieve more opinionative information. We proposed a new method to select opinion bearing terms for query expansion based on a chi-squared test. This was combined in a linear weighting scheme with relevant feedback terms extracted from different search engines and *wikipedia*. This new combined query expansion showed a significant improvement over standard query expansion.

References

[1] S. Abney. Parsing by Chunks. In R. Berwick, S. Abney, and C. Tenny, editors, *Principle-Based Parsing*, Dordrecht, 1991. Kluwer Academic Publishers.

[2] J. Arguello, J. Elsas, J. Callan, , and J. Carbonell. Document representation and query expansion models for blog recommendation. In *ICWSM, International AAAI Conference in Weblogs and Social Media Proceedings*, 2008.

[3] E. Brill. A Simple Rule-based Part of Speech Tagger. In *Third Conference on Applied Natural language processing (ANL) Proceedings*, Trento, Italy, 1992.

[4] B. Ernsting, W. Weerkamp, and M. Rijke. Language modeling approaches to blog post and feed finding. In *Proceedings of the 16th TREC*, 2007.

[5] S. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373, 2004.

[6] S. Kim and E. Hovy. Identifying opinion holders for question answering in opinion texts. In *Proceedings of AAAI 2005 Workshop on Question Answering in Restricted Domains*, 2005.

[7] C. Macdonald and I. Ounis. The trec blogs06 collection: Creating and analysing a blog test collection. Technical report series, University of Glasgow, UK, 2006.

[8] R. Mukras. A comparison of machine learning techniques applied to sentiment classification. Master's thesis, University of Sussex, Falmer, Brighton, 2004.

[9] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL*, pages 271–278, Barcelona, ES, 2004.

[10] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *EMNLP Conference Proceedings*, 2002.

[11] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *EMNLP International Conference Proceedings*, 2003.

[12] E. Riloff, J. Wiebe, and T. Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Natural Language Learning International Conference Proceedings*, pages 25–32, 2003.

[13] V. Stoyanov, C. Cardie, and J. Wiebe. Multi-perspective question answering using the opqa corpus. In *Proceedings of HLT/EMNLP*, 2005.

[14] J. Wiebe. Learning subjective adjectives from corpora. In *Artificial Intelligence National Conference Proceedings*, pages 735–740, 2000.

[15] J. Wiebe, E. Breck, C. Buckly, C. Cardie, P. Davis, B. Fraser, D. Litman, D. Pierce, E. Riloff, and T. Wilson. Multi-perspective question answering. In *Final report of ARDA NRRC Summer Workshop*, 2002.

[16] J. Wiebe, T. Wilson, and M. Bell. Identify collocations for recognizing opinions. In *ACL/EACL Workshop on Collocation Proceedings*, 2001.

[17] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of HLT/EMNLP*, pages 129–136, 2003.

[18] W. Zhang and C. Yu. Uic at trec 2006 blog track. In *Proceedings of the 15th TREC*, 2006.

[19] W. Zhang and C. Yu. Uic at trec 2007 blog track. In *Proceedings of the 16th TREC*, 2007.