WIDIT in TREC-2006 Blog track

Kiduk Yang, Ning Yu, Alejandro Valerio, Hui Zhang
School of Library and Information Science, Indiana University, Bloomington, Indiana 47405, U.S.A.
{kiyang, nyu, avalerio, hz3}@indiana.edu

1. INTRODUCTION

Web Information Discovery Integrated Tool (WIDIT) Laboratory at the Indiana University School of Library and Information Science participated in the Blog track's opinion task in TREC-2006. The goal of opinion task is to "uncover the public sentiment towards a given entity/target", which involves not only retrieving topically relevant blogs but also identifying those that contain opinions about the target. To further complicate the matter, the blog test collection contains considerable amount of noise, such as blogs with non-English content and non-blog content (e.g., advertisement, navigational text), which may misdirect retrieval systems.

Based on our hypothesis that noise reduction (e.g., exclusion of non-English blogs, navigational text) will improve both on-topic and opinion retrieval performances, we explored various noise reduction approaches that can effectively eliminate the noise in blog data without inadvertently excluding valid content. After creating two separate indexes (with and without noise) to assess the noise reduction effect, we tackled the opinion blog retrieval task by breaking it down to two sequential subtasks: on-topic retrieval followed by opinion classification. Our opinion retrieval approach was to first apply traditional IR methods to retrieve on-topic blogs, and then boost the ranks of opinionated blogs based on opinion scores generated by opinion assessment methods. Our opinion module consists of *Opinion Term Module*, which identify opinions based on the frequency of opinion terms (i.e., terms that only occur frequently in opinion blogs), *Rare Term Module*, which uses uncommon/rare terms (e.g., "sooo good") for opinion classification, *IU Module*, which uses IU (I and you) collocations, and *Adjective-Verb Module*, which uses computational linguistics' distribution similarity approach to learn the subjective language from training data.

2. RESEARCH QUESTIONS

In the Blog track, we investigated the following questions:

- Does noise reduction (e.g., exclusion of non-English blogs, navigational text) improve blog retrieval performance?
- How can on-topic retrieval system be optimized to address the challenges of short queries typical in blog retrieval?
- What are the evidences of subjectiveness/opinion and how can they be leveraged to retrieve opinionated blogs?

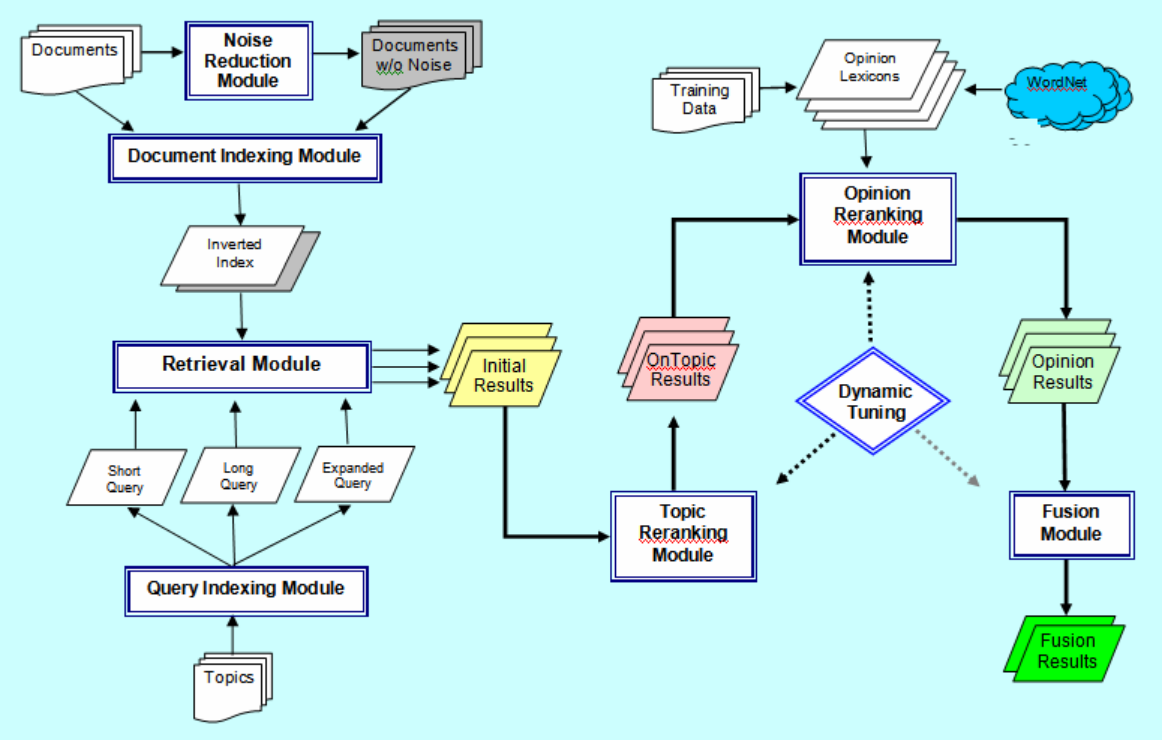
Although noise in data is generally regarded as bad since it dilutes and scrambles the true content to hinder processing and communication of information, it is not clear whether blog-type noise will affect retrieval in any significant manner. In addition to inflating the collection size, noise can mask the true document length thereby misguiding the document-length normalization factor incorporated in term weighting or cosine similarity. In order to assess the effects of noise in blog data, we created two separate indexes: one with noise reduction (i.e., without noise), and the other without noise reduction (i.e., with noise).

In addition to the noise effect question, we focused on the question of how to adapt the existing WIDIT topical retrieval system for opinion retrieval task. The intuitive answer was to first apply existing system to retrieve blogs about a target (i.e., on-topic retrieval), optimize on-

topic retrieval to address the challenges of short queries, and then identify opinion blogs by leveraging evidences of subjectiveness/opinion (i.e., opinion identification). Two key research questions at this point are how to optimize on-topic retrieval, and a compound question of what the evidences of opinion are and how they can be leveraged to retrieve opinionated blogs. As for the opinion identification, we considered the following three 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*").

Figure 1. WIDIT Blog Opinion Retrieval System Architecture



3. METHODOLOGY

WIDIT approach to blog opinion retrieval task consisted of three main steps: initial retrieval, on-topic retrieval optimization, and opinion identification. Initial retrieval was executed using the standard WIDIT retrieval method, on-topic retrieval optimization was done by a post-retrieval reranking approach that leveraged multiple topic-related factors, and opinion identification was accomplished by a fusion of four opinion modules that leveraged multiple sources of opinion evidence. To assess the effect of noise on retrieval performance, we explored various noise reduction methods with which to exclude non-English blogs and non-blog contents from the collection. The overview of WIDIT blog opinion retrieval system is shown in Figure 1.

3.1. Noise Reduction

To effectively eliminate the noise in blog data without inadvertently excluding valid content, we constructed *Non-English Blog Identification (NBI)* module that identifies non-English blogs for exclusion, and *Blog Noise Elimination (BNE)* module that excludes non-blog content portion of the blog. *NBI* leverages the characteristics of non-English (NE) blogs, which contain a large proportion of NE terms, and/or high frequency of NE stopwords. *NBI* heuristic, which scores documents based on NE content density and frequencies of stopwords (both English and non-English), was tuned by iteratively examining the NE blog clusters identified by the module to find false positives and adjusting the NE threshold until no false positives were found. *BNE* module, which uses markup tags to differentiate blog content (e.g., post, comments, etc.) from non-blog content (e.g., scripts, style texts, forms, sidebar, navigation, profile, advertisement, header, footer, etc.), was constructed by examining all unique markup tags in the blog collection to identify patterns to be captured by regular expressions.

3.2. Initial Retrieval

The initial retrieval is executed by the WIDIT retrieval engine, which consists of document/query indexing and retrieval module. After removing markup tags and stopwords, WIDIT's indexing modules applies a modified version of the simple plural remover [7].[1] The stopwords consisted of non-meaningful words such as words in a standard stopword list, non-alphabetical words, words consisting of more than 25 or less than 3 characters, and words that contain 3 or more repeated characters. Hyphenated words were split into parts before applying the stopword exclusion, and acronyms and abbreviations were kept as index terms[2].

In order to enable incremental indexing as well as to scale up to large collections, WIDIT indexes the document collection in fixed-size subcolllections, which are searched in parallel. The whole collection term statistics, derived after the creation of the subcollections, are used in subcollection retrievals so that subcollection retrieval results can simply be merged without any need for retrieval score normalizations.

Query indexing module includes query expansion submodules that identify nouns and noun phrases, expand acronyms and abbreviations, and extract non-relevant portion of topic descriptions with which to formulate various expanded versions of the query.

The retrieval module implements both Vector Space Model (VSM) using the SMART length-normalized term weights and the probabilistic model using the Okapi BM25 formula. For the VSM implementation, SMART *Lnu* weights with the slope of 0.3 are used for document terms [3], and SMART *ltc* weights [2] are used for query terms. *Lnu* weights attempt to match the probability of retrieval given a document length with the probability of relevance given that length [15]. The simplified version of the Okapi BM25 relevance scoring formula [13] is used to implement the probabilistic model.

3.3. On-topic Retrieval Optimization

In order to optimize topical retrieval performance in top ranks, the initial retrieval results are reranked based on a set of topic-related reranking factors. The topic reranking factors used are *Exact Match*, which is the frequency of exact query string occurrence in document normalized by document length, *Proximity Match*, which is the length-normalized frequency of padded[3] query string occurrence, *Noun Phrase Match*, which is the length-normalized frequency of query noun

[1] The simple plural remover was chosen to speed up indexing time and to minimize the overstemming effect of more aggressive stemmers.
[2] Acronym and abbreviation identification was based on simple pattern matching of punctuations and capitalizations.
[3] "Padded" query string is a query string with up to *k* number of words in between query words.

phrases occurrence, and *Non-Rel Match*,[4] which is the length-normalized frequency of non-relevant nouns and noun phrase occurrence. The on-topic reranking method consists of following three steps:

(1) Compute topic reranking scores for each of top N results.

(2) Categorize the top N results into reranking groups designed to preserve initial ranking while appropriate rank-boosting for a given combination of reranking factors.

(3) Boost the rank of documents using reranking scores within groups.

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).

3.4. Opinion Identification

Opinion identification is accomplished by combining the four opinion modules that leverage various evidences of opinion (e.g, Opinion Lexicon, Opinion Collocation, Opinion Morphology). The modules are *Opinion Term Module*, which identify opinions based on the frequency of opinion terms (i.e., terms that only occur frequently in opinion blogs), *Rare Term Module*, which uses uncommon/rare terms (e.g., "sooo good") for opinion classification, *IU Module*, which uses IU (I and you) collocations, and *Adjective-Verb Module*, which uses computational linguistics' distribution similarity approach to learn the subjective language from training data. Opinion modules require opinion lexicons, which are extracted from training data. We constructed 20 training topics from BlogPulse (http://www.blogpulse.com/) and Technorati search (http://www.technorati.com/) archives and manually evaluated the search results of the training topics to generate the training data set of 700 blogs. The application of opinion modules is similar to on-topic retrieval optimization in that opinion scores generated by modules act as opinion reranking factors to boost the ranks of opinionated blogs in the topic-reranked results.

3.4.1. Opinion Term Module

The basic idea behind the *Opinion Term Module* (OTM) is to identify opinion blogs based on the frequency of opinion terms, which are terms that only occur frequently in opinion blogs. OTM computes opinion score using an OT lexicon, which we created by extracting terms from positive training data using information gain, excluding terms appearing in negative training data, and manually selecting a set of opinion terms. Two OTM scores are generated: document-length normalized frequency of OT terms in document and OT terms near query string in document.

3.4.2. Rare Term Module

Rare Term Module (RTM) is derived from the hypothesis that people become creative when expressing opinions and tend to use uncommon/rare terms (e.g., "sooo good"). Thus, we extracted low frequency terms from positive training data, removed dictionary terms, and examined them to construct a RT lexicon and regular expressions that will identify creative term patterns used in opinion blogs. Two RT scores similar to OT scores are computed.

3.4.3. IU Module

IU Module (IUM) is based on the observation that pronouns such as 'I' and 'you' appear very frequently in opinion blogs. For IU lexicon construction, we compiled a list of IU (I and you)

[4] Non-rel Match is used to suppress document instead of boosting.

collocations from training data (e.g., 'I believe', 'my assessment', 'good for you', etc.). IUM counts the frequency of "padded" IU collocations within sentence boundary to compute two IUM scores similar to OTM and RTM.

3.4.4. Adjective-Verb Module

The hypothesis underlying Adjective-Verb module (AVM) is similar to OTM in that it assumes high frequency of opinion terms in opinion blogs. In addition to restricting opinion terms to verbs and adjectives, AVM differs from OTM in its lexicon construction by using computational linguistics' distribution similarity approach that attempts to learn the subjective language from training data rather than shallow linguistic approaches of other opinion modules. The Adjective/Verb component uses the density of potentially subjective elements (PSE) to determine the subjectivity of blog posts. It assumes that a post with a high concentration of subjective adjectives and verbs must be opinionated. These parts of speech are the ones that better reveal the author's intention by using attributes ("good", "bad", "ugly") or expressing reactions to ideas or objects ("hate", "love", "disgust"). The idea was evaluated by Wiebe et al. [17] with successful results and their algorithm was the starting point for the design of the component.

The component relies heavily on the elements of the PSE set, so their selection is a key process that must be done carefully. Ideally, the PSE set should be broad so that the wide variety of terms used to describe opinion is captured, but at the same time should not include ambiguous terms that may lead to false positives. For this purpose, an initial PSE set of subjective terms is manually collected. The seed set is then expanded, first by gathering related terms from several lexical references, and second by finding terms that co-occur with PSEs in opinionated posts. Next, the set of candidate PSE is refined by verifying its classification performance against a validation set and removing the elements that lead to misclassifications. The PSE set is cleaned up manually at the end of the process and also at several points between the execution steps.

3.5. Fusion

The fusion module combines the multiple sets of search results after retrieval time. In addition to two of the most common fusion formulas, *Similarity Merge* [6, 7] and *Weighted Sum* [1, 15], WIDIT employs variations of the weighted sum formula. The similarity merge formula multiplies the sum of fusion component scores for a document by the number of fusion components that retrieved the document (i.e. overlap), based on the assumption that documents with higher overlap are more likely to be relevant. Instead of relying on overlap, the weighted sum formula sums fusion component scores weighted with the relative contributions of the fusion components that retrieved them, which is typically estimated based on training data. Both formulas compute the fusion score of a document by a linear combination of fusion component scores.

In our earlier study [20], similarity merge approach proved ineffective when combining content- and link-based results, so we used a variation of the weighted sum fusion formula that sums the normalized system scores multiplied by system contribution weights [19]. One of the main challenges in using the weighted fusion formula lies in determination of the optimum weights for each system. In order to optimize the fusion weights, WIDIT engages in a static tuning process, where various weight combinations are evaluated with the training data in a stepwise fashion.

3.6. Dynamic Tuning

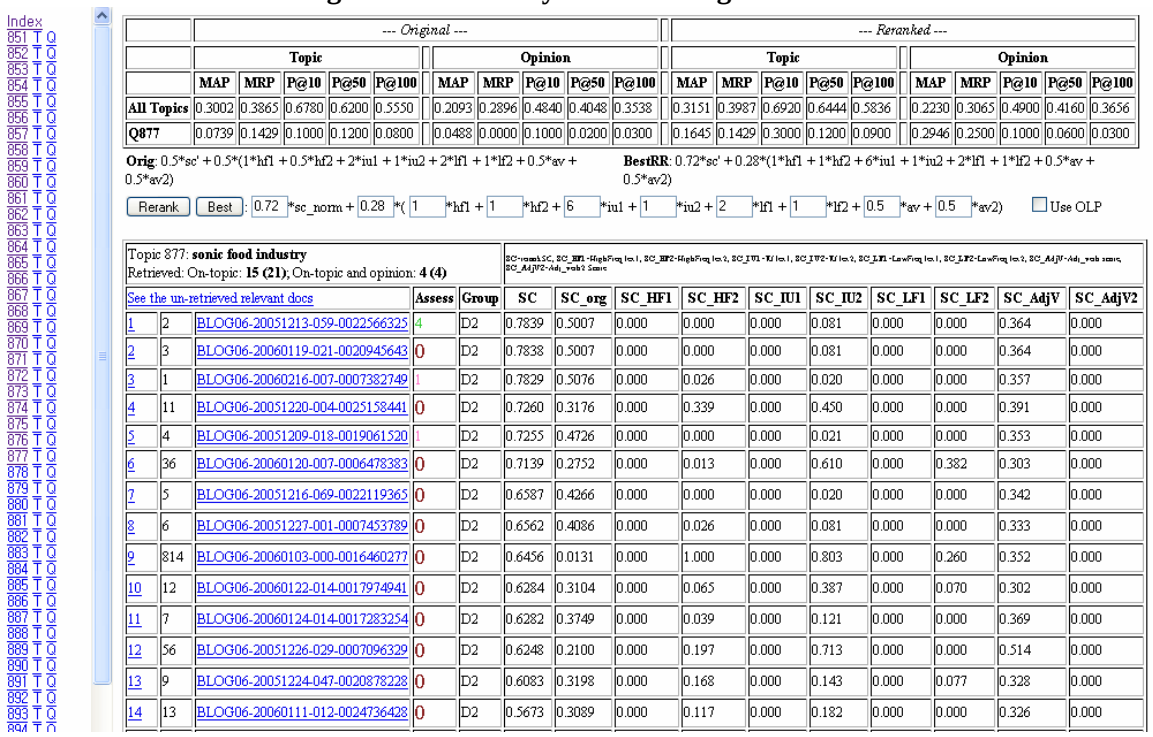
Both topic and opinion reranking involve combination of multiple reranking factors as can be seen in the generalized reranking formula below:

$$RS = \alpha * NS_{orig} + \beta * \sum (w_i * NS_i) \qquad (1)$$

In formula (1), NS_{orig} is the normalized original score, NS_{orig} is the normalized score of reranking fator i, w_i is the weight of reranking factor i, α is the weight of original score, and β is the weight of the overall reranking score.

To optimize the reranking formulas, which involve determination of optimum reranking factor weights (w_i), we implemented *Dynamic Tuning* (Figure 2), which is a bio-feedback like mechanism that displays effects of tuning parameter changes in real time to guide human to find the local optimum. The key idea of dynamic tuning, which is to combine the human intelligence, especially pattern recognition ability, with the computational power of the machine, is implemented in a Web application that allows 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.

Figure 2. WIDIT Dynamic Tuning Interface



4. RESULTS

After the official submission, we conducted post-submission experiments that involved optimization of reranking and tuning modules using relevance data as well as overall system refinements. Among numerous system parameters at play, we examined the effects of following independent variables on retrieval performance using the post-submission results: noise reduction, query length, topic reranking, opinion reranking, dynamic tuning, and fusion. The TREC official results of top 5 groups are displayed below (Table 1).

Table 1. Official TREC blog opinion results of top 5 systems

Group	MAP	MRP	P@10
Indiana University	**0.2052**	**0.2881**	0.468
Univ. of Maryland	0.1887	0.2421	0.378
Univ. of Illinois at Chicago	0.1885	0.2771	**0.512**
Univ. of Amsterdam	0.1795	0.2771	0.464
Univ. of California, Santa Cruz	0.1549	0.2355	0.438

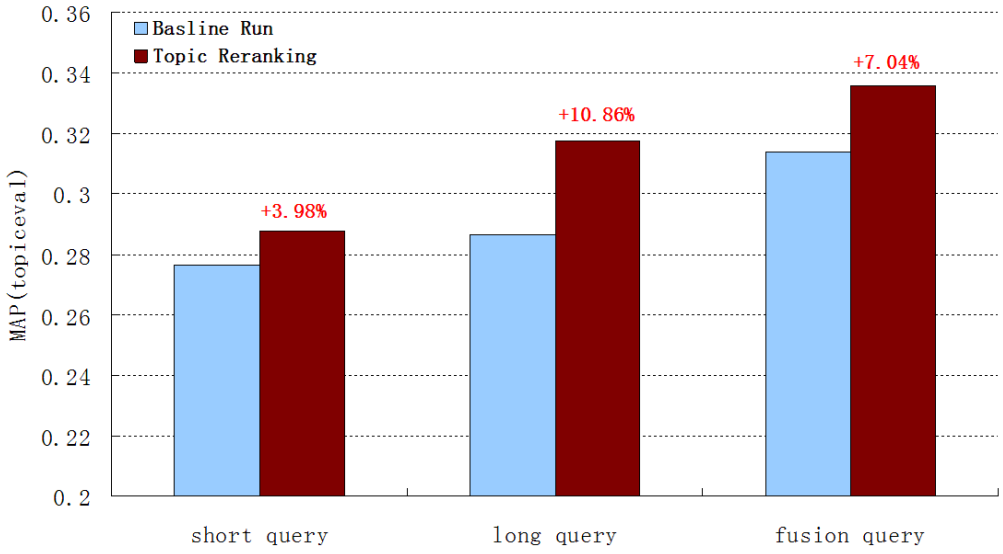
4.1. Noise Reduction Effect

Contrary to the assumption that noise reduction will improve retrieval performance, our noise reduction approach only slightly increased P@10, while generally decreasing MAP and MRP. Upon examination of the excluded data, we discovered that our noise reduction module eliminated some true blog content despite our attempts to be conservative. Although the effect of noise reduction on retrieval is inconclusive, we still believe that noise reduction is a good idea as evidenced by the increase in high precision even with the faulty implementation.

4.2. Query Length Effect

It is well-know fact in information retrieval community that longer queries in general will produce better retrieval result. This was shown to hold true for blog opinion retrieval as well. Figure 3 shows consistently superior performances of longer queries in all phases of retrieval (i.e, initial retrieval, topic-reranking, topic reranking with dynamic tuning, opinion reranking, opinion reranking with dynamic tuning), and by both the topical and opinion performance evaluation. One exception occurs with baseline topic retrieval performance of the long query (title, description, narrative), which is worse than that of the short query. This may be due to introduction of noise in the long query, which is consistent with our past work that found some long queries to be harmful for finding specific targets due to introduction of noise [20]. When the same results are evaluated with opinion relevance (lower three line in Figure 3), however, the long query performs same as the short query. This suggests that the long query may contain description of opinions that helps finding opinion blogs while retrieving non-topical blogs at the same time. This anomaly is corrected by reranking strategy that uses combination of key evidences to boost the ranks of blogs likely to be relevant.

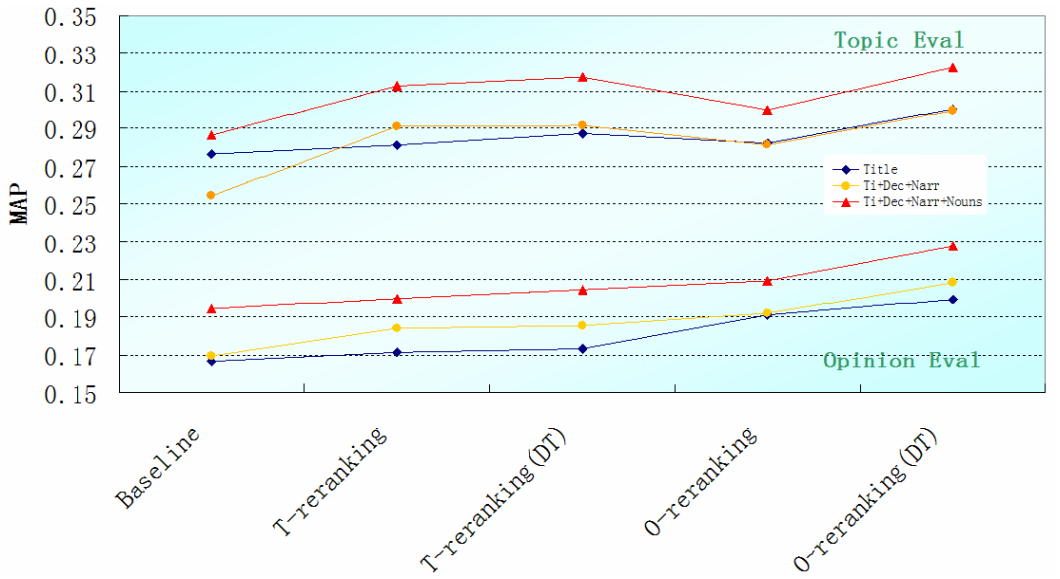
Figure 3. Query Length Effect



4.3. Topic Reranking Effect

The effect of topic reranking on initial retrieval is shown in Figure 4. The gain in topic retrieval performance by topic reranking is marginal for the short query (4%) but over 10% improvement for the long query. This is understandable since topic reranking factors capitalize on topical evidence, which the short queries have little of.

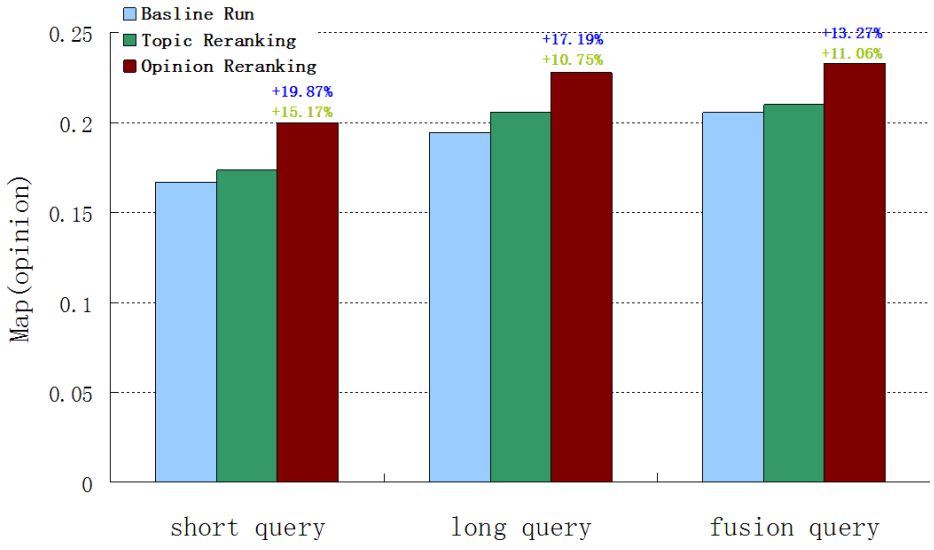
Figure 4. Topic Reranking Effect



4.4. Opinion Reranking Effect

Figure 5 displays the marked effects of opinion reranking. For the short query, opinion reranking improves the performance of topic reranked results by 15% (20% over baseline) and for the long query, 11% improvement (17% over baseline). It clearly demonstrates the effectiveness of WIDIT's opinion reranking approach.

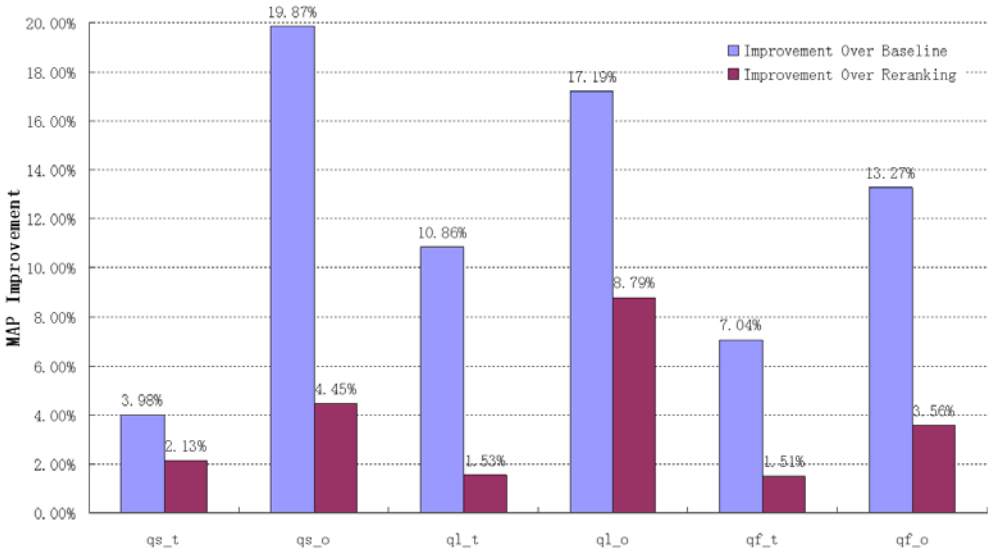
Figure 5. Opinion Reranking Effect



4.5. Dynamic Tuning Effect

The effect of dynamic tuning is shown in Figure 6. Since the blue bars show improvements over baseline that contain the reranking effect, the isolated effect of dynamic tuning turns out to be only marginal (4.5% for short query and 9% for long query). We suspect this is partially influenced by reranking effect that took the system performance towards the ceiling and partially by the mostly linear nature of tuned formulas that require more rule-based intervention to approach the optimum solution space.

Figure 6. Dynamic Tuning Effect



4.6. Fusion Effect

As we have repeatedly found in previous research [20, 21, 22], the fusion approach is shown to be quite beneficial (Table 2). Fusion, which shows the best overall performance of all system combinations, improves performance by 20% over best baseline non-fusion result.

Table 2. Opinion MAP of best baseline and fusion results

	QShort	QLong	Fusion
Baseline	.1666	.1943	.2057
Reranked			
- no Tuning	.1912	.2093	.2250
- DTuning	.1997	.2277	**.2230**

5. CONCLUSION

WIDIT's fusion approach of combining multiple sources of evidence and multiple methods worked well for TREC's blog opinion retrieval task. Topic and opinion reranking, as well as fusion all contributed to improving retrieval performance, and the compound effect of all three resulted in the best overall performance. Although opinion retrieval posed non-trivial challenges, stepwise approach of initial retrieval, on topic retrieval optimization, and opinion identification proved to be an effective solution.

REFERENCES

[1] Bartell, B. T., Cottrell, G. W., & Belew, R. K. (1994). Automatic combination of multiple ranked retrieval systems. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

[2] Buckley, C., Salton, G., & Allan, J., & Singhal, A. (1995). Automatic query expansion using SMART: TREC 3. *Proceeding of the 3rd Text Rerieval Conference (TREC-3)*, 1-19.

[3] Buckley, C., Singhal, A., & Mitra, M. (1997). Using query zoning and correlation within SMART: TREC 5. *Proceeding of the 5th Text REtrieval Conference (TREC-5)*, 105-118.

[4] Chklovski, T. (2006). Deriving quantitative overviews of free text assessments on the web. In IUI '06: *Proceedings of the 11th international conference on Intelligent User Interfaces*, New York, NY, USA, pp. 155–162. ACM Press.

[5] Efron, M. (2004). The liberal media and right-wing conspiracies: using cocitation information to estimate political orientation in web documents. *Proceedings of the thirteenth ACM international conference on Information and Knowledge Management*, 390–398.

[6] Fox, E. A., & Shaw, J. A. (1995). Combination of multiple searches. *Proceeding of the3rd Text Rerieval Conference (TREC-3)*, 105-108.

[7] Frakes, W. B., & Baeza-Yates, R. (Eds.). (1992). *Information retrieval: Data structures & algorithms*. Englewood Cliffs, NJ: Prentice Hall.

[8] Herring, S. C., I. Kouper, J. C. Paolillo, L. A. Scheidt, M. Tyworth, P. Welsch, E. Wright, and N. Yu (2005). Conversations in the blogosphere: An analysis "from the bottom up". *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)*.

[9] Hu, M. and Liu, B. (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.

[10] Lee, J. H. (1997). Analyses of multiple evidence combination. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 267-276.

[11] Liu, B., 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.

[12] Mishne, G. 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 (WWW2006)*.

[13] Robertson, S. E. & Walker, S. (1994). Some simple approximations to the 2-Poisson model for probabilistic weighted retrieval. *Proceedings of the* 17th *ACM SIGIR Conference on Research and Development in Information Retrieval*, 232-241.

[14]Savoy, J., & Picard, J. (1998). Report on the TREC-8 Experiment: Searching on the Web and in Distributed Collections. *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, 229-240.

[15] Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 21-29.

[16] Thompson. P. (1990). A combination of expert opinion approach to probabilistic information retrieval, part 1: The conceptual model. *Information Processing & Management*, 26(3), 371-382.

[17] Wiebe, J., T. Wilson, R. Bruce, M. Bell, and M. Martin (2004). Learning subjective language. Comput. Linguist. 30 (3), 277–308.

[18] Wilson, T., 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.

[19] Yang, K. (2002a). Combining Text-, Link-, and Classification-based Retrieval Methods to Enhance Information Discovery on the Web. (*Doctoral Dissertation*. University of North Carolina).

[20] Yang, K. (2002b). Combining Text- and Link-based Retrieval Methods for Web IR. *Proceedings of the 10th Text Rerieval Conference (TREC2001)*, 609-618.

[21] Yang, K., & Yu, N. (2005). WIDIT: Fusion-based Approach to Web Search Optimization. *Asian Information Retrieval Symposium 2005*.

[22] Yang, K., Yu, N., Wead, A., La Rowe, G., Li, Y. H., French, C., & Lee, Y (2005). WIDIT in TREC2004 Genomics, HARD, Robust, and Web tracks. *Proceedings of the 13th Text Retrieval Conference (TREC2004).*