Retrieving Opinions from Discussion Forums

Laura Dietz[1], Ziqi Wang[2], Samuel Huston[1] and W. Bruce Croft[1]

[1] Center for Intelligent Information Retrieval
School of Computer Science
University of Massachusetts Amherst
Amherst, MA, 01002, USA
{dietz,sjh,croft}@cs.umass.edu

[2] School of EECS
Peking University
Beijing 100871, China
wangziqi@pku.edu.cn

ABSTRACT

Understanding the landscape of opinions on a given topic or issue is important for policy makers, sociologists, and intelligence analysts. The first step in this process is to retrieve relevant opinions. Discussion forums are potentially a good source of this information, but comes with a unique set of retrieval challenges. In this short paper, we test a range of existing techniques for forum retrieval and develop new retrieval models to differentiate between opinionated and factual forum posts. We are able to demonstrate some significant performance improvements over the baseline retrieval models, demonstrating that this as a promising avenue for further study.

Categories and Subject Descriptors

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

General Terms

Experimentation, Performance

Keywords

Opinion retrieval, social media, relevance model

1. INTRODUCTION

In this paper, we investigate the retrieval of opinions from social media sources. Formally, given an information need, we retrieve documents that contain topically relevant expressions of opinion. In this context, information needs are expressed as longer grammatical queries, for example; *What is causing the real estate crisis in the USA?* Note that there are many factual answers to this question, and many more opinions on these answers. In this paper, we present initial research on this problem and explain its idiosyncrasies.

There are a number of activities that could take advantage of high quality opinion retrieval. Some examples include:

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2507861.

- Political decisions require summaries of arguments, and opinions of various possible policies. The automatic extraction of public opinion from social media facilitates the preparation of the briefing books used to make those decisions.
- Social science research often asks research questions that require the collection and analysis of opinions expressed by various cohorts within society.
- Intelligence agencies could use this type of resource to enable prediction of major societal events, such as the Arab Spring.

Each of these high level tasks requires some post-processing of retrieved documents to extract relevant opinions from documents, such as passage extraction or summarization. In this paper, we focus on the information retrieval techniques, in which the system returns a high quality, ranked set of topically relevant documents containing expressions of opinions. Information extraction methods can then be applied to these documents to produce the final output for the particular task.

Retrieving opinions from discussion forums on particular topics and questions is very different from sentiment analysis of product reviews [13]. Product reviews follow template-like language with a small set of attributes expressing pros and cons, which have been successfully tackled with dictionaries such as SentiWordNet [2]. In contrast, language that expresses opinions about politics, economics, and ethics is much more subtle and diverse, and is hard to detect using dictionary-based approaches, as is demonstrated in this paper. Generally, opinion retrieval caters to two typical use cases, poll and exploration. For opinion polls, the task is to estimate the fraction of users who agree or disagree with an opinion. This paper addresses the exploration use case, where a portfolio of opinionated answers to a question are to be collected.

Retrieval over discussion forums poses challenges across a number of dimensions. The data is frequently agrammatical and the vocabulary is often invented on the spot. A particular problem we encounter involves the disambiguation of pseudonyms for public figures. The culture of quoting in replies leads to an extreme duplication of content. Although most forum posts express opinions, users often quote Wikipedia and news articles in place of an answer. Selection of material is one way of expressing opinion, but we are primarily interested direct expressions of first hand opinions.

Previous studies of social media retrieval have used various techniques to cope with these aspects of social media search.

For example, the thread and conversation structure have been shown to aid retrieval by providing context for short postings [17, 6]. External information sources have also been shown to be able to improve retrieval performance [7]. In addition, a wide variety of techniques have been proposed to cope with the low quality and high degree of repeated content in microblogs [12, 3]. Sentiment analysis, the act of determining both if a document expresses an opinion and its polarity is an active area of research [13, 15]. Akkaya et al. [1] study context-based classifiers that distinguish subjective from objective use of opinionated words. Okamoto et al. [14] investigates methods of incorporating sentiment analysis into sentence retrieval using contextual data. Huang and Croft [9] explore different query expansion methods for opinion retrieval on blog data.

In this short paper, we analyze the effectiveness of some of these techniques for opinion retrieval, leading to a retrieval benchmark for opinion retrieval in forum data.

The remainder of this paper is laid out as follows. We begin with a description of the query models for forum retrieval in Section 2.1 and focus on methods for opinion retrieval in Section 2.2. The discussion forum collection and the test sets of opinion-focused queries is described in Section 3. We detail experimental evaluation in Section 4 before concluding the paper.

2. APPROACHES

2.1 Forum Retrieval

We start by investigating a wide variety of effective retrieval models. Baseline retrieval models include the query likelihood (QL) [16] and the sequential dependence model (SDM) [11]. Further, the sequential dependence model is extended using pseudo relevance feedback (RM3) [10].

Recent research has shown that performing query expansion using data extracted from Wikipedia can dramatically improve performance for a variety of information retrieval problems [7, 18]. We investigate two methods of expanding queries using Wikipedia data. First, we perform pseudo-relevance feedback using RM3 [7, 10] over Wikipedia articles to determine expansion terms (WikiRM3). Second, we use Wikipedia titles and redirection pages to determine synonym sets for query terms (WikiRD) using the method proposed by Xue and Croft [19]. WikiRD candidates are accepted based on the context model score, as proposed by Dang and Croft [5].

2.2 Opinionated Pseudo Relevance Feedback

When searching for people's opinions on a given topic, we want to bias against forum posts that are objectively talking about the topic. In this line of research we mainly care for the valence of the sentiment, independent of its polarity. We evaluate opinionated expansion methods for SDM, as well as in combination with the forum retrieval methods from Section 2.1.

We explore three query-independent expansion methods which have shown good results for blog sentiment retrieval [9]: The first method (Seed) expands with a fixed set of sentiment seed words: "good, nice, excellent, positive, fortunate, correct, superior, bad, nasty, poor, negative, unfortunate, wrong, inferior". The second method (OP) expands queries with terms that are both frequent in the forum data, and present in the external opinionated corpora General Inquirer[1] and Opinion Finder's subjectivity Lexicon.[2] Thirdly, we explore a supervised method (SV) which selects terms from the target collection. Candidate terms are selected from the top k documents for each query in a set of training queries. The terms are weighted to maximize the mean average precision over the training queries. As this method requires training data, we employ 2-fold cross-validation within each test collection.

Next, we explore query-specific expansion methods. We start with a variant on pseudo-relevance feedback methods, where each candidate expansion term is required to be present in a term list or dictionary. We separately evaluate the resources used in Huang's & Croft's work (FiltO) and subjective words in SentiWordNet (FiltS) [2].[3]

Usually, pseudo relevance feedback associates an expansion term t for a given query q with a probabilistic weight $p_C(t|q) = \sum_d p_C(d|q)p(w|d)$, where $p_C(d|q)$ refers to the retrieval probability of the target corpus C. We recognize that SentiWordNet contains many words that indicate subjectivity in product reviews but not so much in ethical discussions. Instead of adding terms to bias towards subjectivity, it is possible to bias against objectivity. In this study, we choose to use news data to discover query-specific objective terms. We study a modified relevance model we call Anti-news Relevance Model (ARM) that requires expansion terms to be in SentiWordNet but also biases against words that are common in a news collection N. We formalize this as a likelihood ratio test and expand terms with weights $w_C^{aRM}(t|q) = \frac{p_C(t|q)}{p(t|N)}$. We also explore a variant (FiltN) that uses the top words selected by ARM, but weights terms according to the standard RM weighting $p_C(t|q)$. Similar to SV, each of these models require training data, we employ 2-fold cross-validation within each test collection

All sentiment expansion methods are intended to be used in a weighted combination with the original query and (non-opinionated) relevance model. All word counts are smoothed with Dirichlet smoothing; all expansion weights are normalized to sum to 1, projecting probabilities and likelihood ratios onto the same range.

3. CORPUS, QUERIES, AND JUDGMENTS

To investigate opinion retrieval, we use a large collection of forum data, collected by the Linguistic Data Consortium (LDC).[4] The set of forums crawled discuss a range of issues, with topics ranging from political and news events to medical subjects. In this paper, we use a subset consisting of approximately 262,000 threads, containing almost 5.5 million posts. The subset contains 591 million terms, with a vocabulary of 1.1 million unique terms.

In the evaluation of opinion retrieval models, we use two query test sets. The first is compiled and judged by LDC to which we refer to as "P1-Eval". This test set contains 146 questions. A total of 2172 relevant passage-level judgments are identified by LDC from a pool of rankings that did not include our methods. Since our task is to retrieve documents for passage extraction, we use the best passage judgment as a measure of document relevance.

[1] http://www.wjh.harvard.edu/~inquirer
[2] http://www.cs.pitt.edu/mpqa
[3] Single word expressions with PosScore + NegScore ≥ 0.25.
[4] http://www.ldc.upenn.edu/

Table 1: Retrieval performance measured using P@10, and nDCG@10. Significant improvement with respect to the SDM baseline is marked^{+}.

	Development		P1-Eval	
	P@10	nDCG@10	P@10	nDCG@10
QL	0.707	0.547	0.215	0.271
SDM	0.736	0.568	**0.232**	**0.281**
SDM+RM3	0.748	0.577	0.229	0.275
SDM+WikiRD	0.760	0.584	0.21	0.25
SDM+WikiRM3	0.717	0.548	0.230	0.28
SDM+RM3+WikiRD	**0.774**$^{+}$	**0.588**	0.216	0.256

We independently compile an additional development test set of 42 opinionated queries from questions posed on Yahoo! Answers[5] that have elicited opinionated responses on Yahoo! Answers. The associated user posted answers are not used in the this study. All queries were modified to be grammatically correct questions and are available online.[6] For example:

Q1 *Who is responsible for the deteriorating economy?*
Q2 *What can be done about Somali Pirates?*

Relevance judgments on the forum data are created by us using a pooling method of several supervised and unsupervised retrieval methods including approaches in Section 2.1. Below are some excerpts from relevant forum posts for the above queries.

Q1 *... the market didn't go down because of the Democrats. It went down because the housing market was in the biggest bubble of all time ...*
Q2 *... Secondly, allow me to put my position across. The argument that these young men in some way deserved what they got because they were sailing through a notoriously risky stretch of ocean no longer applies in this case,...*

For each development query, a pool of the top ten returned posts from each retrieval model were judged for three-valued topical relevance. Across the 42 queries, 2099 forum posts were identified as relevant or partially relevant. While judging a pool of only highly ranked posts yields a relatively shallow pool of judgments, it is enough to differentiate performance between some of the retrieval methods. In accordance with this annotation process, we focus our analysis on P@10 and nDCG@10.

In addition to the forum data, information extracted from several external data sources are used to complement retrieval models. A dump of almost 10 million articles from Wikipedia, downloaded in June 2012, is used in some query expansion techniques. As a large, recent collection of news articles, we use a collection of around 134 million news documents, originally collected for use in the TREC 2012 Knowledge-base Acceleration Track [8].

4. EXPERIMENTS

All retrieval models were implemented using the Indri Search Engine [4]. Results based on the pool of judgments of the development set from all retrieval experiments and official judgments compiled by LDC on P1-Eval are shown in Table 1. Statistical differences in this work are tested using a two-tailed paired t-test with level $\alpha = 5\%$.

[5]http://answers.yahoo.com/
[6]http://www.ciir.cs.umass.edu/~dietz/forum-opinion/

Table 2: Retrieval performance among opinionated expansion methods. Significance over SDM baseline is marked $^{+}$.

	Development		P1-Eval	
	P@10	nDCG@10	P@10	nDCG@10
SDM	0.736	0.568	**0.232**	**0.281**
SDM+Seed	0.712	0.561	0.228	0.278
SDM+OP	0.717	0.558	0.23	0.278
SDM+SV	0.719	0.562	0.225	0.274
SDM+FiltO	0.733	0.574	0.224	0.272
SDM+FiltS	**0.74**	**0.582**	0.215	0.266
SDM+ARM	0.724	0.562	0.225	0.275
SDM+FiltN	0.736	0.578	0.225	0.276
SDM+FiltN+RM3	0.707	0.561	**0.229**	0.273
SDM+FiltS+RM3	0.757	0.580	0.227	**0.275**
SDM+WikiRD+FiltS+RM3	**0.783**$^{+}$	**0.597**	0.213	0.255

For the Development query set, we can see that for P@10 and nDCG@10 the combination of the SDM, RM3, and WikiRD is the most effective model, significantly outperforming SDM on P@10. Metrics on the P1-Eval query set yield low numbers, because the majority of the top 10 retrieved documents are unjudged. Therefore we also observe only very small differences between each of the models. However, we can clearly see that WikiRD degrades average performance over the P1-Eval query set, inspection of per-query results reveals that this technique dramatically harms the performance of just 10% of query set, relative to SDM.

Results on opinionated expansion methods are presented in Table 2. Mixing weights between original query, RM, and opinionated expansion model are learned together with the numbers of expansion documents and terms via 2-fold cross-validation. On both data sets we observe a consistent improvement with opinionated filtering of relevance model expansion terms. We note that combining SDM with each query independent expansion model deteriorates performance, relative to SDM alone. The query-dependent expansion techniques, however, show some small, but promising improvements over the SDM baseline.

We show detailed results and expansion terms for the test question, *What is the republicans solution for the healthcare system?*, in Table 3. We display cumulative performance increases when adding RM, FiltS, and WikiRD to SDM. We observe this improvement in more than half of the development queries, and this improvment is further reflected in as a significant improvement over the P@10 metric.

Inspecting the lists of expansion terms from the different methods for this example query, we confirm that FiltO selects very opinionated expansion terms, and ARM identifies forum-typical expressions. FiltS includes many topical words. However, even in combination with opinion ag-

Table 3: Expansion terms and performance for question "What is the republicans solution for the healthcare system?"

RM	WikiRM	Seed	OP	SV	FiltO	FiltS	ARM	FiltN
america	carolina	good	even	drive	problem	health	malpractice	health
health	allscript	bad	need	even	oppose	govern	unfairness	care
public	ge	nice	point	force	matter	care	uninsured	pay
insurance	certify	nasty	mean	mean	point	pay	want	make
make	data	excellent	try	too	bankrupt	make	out	state

	P@10	NDGC@10
SDM	0.7	0.63
+ RM3	0.8	0.66
+ FiltS	**0.9**	0.68
+WikiRD	0.8	**0.80**

nostic RM3, FiltS achieves better performance than FiltO and ARM on the complete test set. We observe that query-specific expansion techniques perform consistently better than query-independent expansion techniques.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we present initial research on retrieving opinions in discussion forum data. We encountered several interesting problems relating to the nature of social media. In particular, we found several issues relating to the noisy nature of forum data, duplication of content, and verbatim replication of news and Wikipedia articles.

We observe that many successful retrieval models for adhoc web retrieval do not outperform baseline retrieval models for this task, indicating that this problem deserves attention. We achieved good results with a filter-approach based on a list of sentiment words from product reviews, although they contain many words of topical relevance. Methods for subjectivity word sense disambiguation [1] may help distinguish those cases. However, we note that even in combination with RM3 models, expansion using mix of topical and opinonated words (FiltS) achieves better performance than filtering with strictly opinionated words (FiltO).

This initial research demonstrates that opinion retrieval remains an open problem for information retrieval. Future work will include further study on how to promote "good" and suppress "bad" expansion terms from external corpora, such as news and Wikipedia articles.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part under subcontract #19-000208 from SRI International, prime contractor to DARPA contract #HR0011-12-C-0016, and in part by NSF grant #CNS-0934322 Any opinions, findings expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

[1] Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. Subjectivity word sense disambiguation. In *Proc. of the EMNLP*, pages 190–199, 2009.

[2] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. 2010.

[3] Jaeho Choi, W. Bruce Croft, and Jin Young Kim. Quality models for microblog retrieval. In *Proc. of the 21st ACM CIKM*, pages 1834–1838, 2012.

[4] W. Bruce Croft and Jamie Callan. The Lemur Project. http://www.lemurproject.org/, 2001-2012. Lemur Toolkit, Indri, Galago, ClueWeb09.

[5] Van Dang and W. Bruce Croft. Query reformulation using anchor text. In *Proc. of the third ACM WSDM*, pages 41–50, 2010.

[6] Jonathan L. Elsas and Jaime G. Carbonell. It pays to be picky: an evaluation of thread retrieval in online forums. In *Proc. of the 32nd ACM SIGIR*, pages 714–715, 2009.

[7] Jonathan L. Elsas, Jaime Arguello, Jamie Callan, and Jaime G. Carbonell. Retrieval and feedback models for blog feed search. In *Proc. of the 31st ACM SIGIR*, pages 347–354, 2008.

[8] J. R. Frank, M. Kleiman-Weiner, D. A. Roberts, F. Niu, C. Zhang, C. Re, and I. Soboroff. Building an Entity-Centric stream filtering test collection for TREC 2012. In *Proc. of the TREC*, 2012.

[9] Xuanjing Huang and W. Bruce Croft. A unified relevance model for opinion retrieval. In *Proc. of the 18th ACM CIKM*, 2009.

[10] Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *Proc. of the 24th ACM SIGIR*, pages 120–127, 2001.

[11] Donald Metzler and W. Bruce Croft. A Markov random field model for term dependencies. In *Proc. of the 28th ACM SIGIR*, pages 472–479, 2005.

[12] Donald Metzler, Congxing Cai, and Eduard Hovy. Structured event retrieval over microblog archives. In *Proc. of the 2012 NAACL HLT*, pages 646–655, 2012.

[13] Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: capturing favorability using natural language processing. In *Proc. of the 2nd K-CAP*, pages 70–77, 2003.

[14] Takayoshi Okamoto, Tetsuya Honda, and Koji Eguchi. Locally contextualized smoothing of language models for sentiment sentence retrieval. In *Proc. of the 1st CIKM workshop TSA*, pages 73–80, 2009.

[15] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2): 1–135, 2008.

[16] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proc. of the 21st ACM SIGIR*, pages 275–281, 1998.

[17] Jangwon Seo, W. Bruce Croft, and David A. Smith. Online community search using conversational structures. *Inf. Retr.*, 14(6):547–571, December 2011.

[18] Yang Xu, Gareth J.F. Jones, and Bin Wang. Query dependent pseudo-relevance feedback based on Wikipedia. In *Proc. of the 32nd ACM SIGIR*, pages 59–66, 2009.

[19] Xiaobing Xue and W. Bruce Croft. Generating reformulation trees for complex queries. In *Proc. of the 35th ACM SIGIR*, pages 525–534, 2012.