Reproducing and Generalizing Semantic Term Matching in Axiomatic Information Retrieval

Peilin Yang[1] and Jimmy Lin[2(✉)]

[1] Ontario, Canada
[2] David R. Cheriton School of Computer Science, University of Waterloo, Ontario, Canada
jimmylin@uwaterloo.ca

Abstract. In the framework of axiomatic information retrieval, the semantic term matching technique proposed by Fang and Zhai in SIGIR 2006 has been shown to be effective in addressing the vocabulary mismatch problem, with experimental evidence provided from newswire collections. This paper reproduces and generalizes these results in Anserini, an open-source IR toolkit built on Lucene. In addition to making an implementation of axiomatic semantic term matching available on a widely-used open-source platform, we describe a series of experiments that help researchers and practitioners better understand its behavior across a number of test collections spanning newswire, web, and microblogs. Results show that axiomatic semantic term matching can be applied on top of different base retrieval models, and that its effectiveness varies across different document genres, each requiring different parameter settings for optimal effectiveness.

Keywords: Axiomatic retrieval · Query expansion

1 Introduction

The *vocabulary mismatch* problem is one of the most fundamental challenges in information retrieval. Frequently, query terms expressing an information need differ from those used by authors of relevant documents. Retrieval models based on exact term matches, which include instances from the probabilistic retrieval family, language modeling framework, and many others, have difficulty with this problem. "Classic" approaches to tackling this challenge include relevance feedback [7], query expansion [8,9], and modeling term relationships using statistical translation [1], while a new generation of neural ranking models offer solutions based on continuous word representations [6]. In this paper, we focus on reproducing and generalizing an alternative approach to addressing the vocabulary mismatch problem in the axiomatic retrieval framework [2]—specifically, the SIGIR 2006 paper of Fang and Zhai [3] (henceforth, FZ for short). The paper showed that semantic term matching can be incorporated into the axiomatic

L. Azzopardi et al. (Eds.): ECIR 2019, LNCS 11437, pp. 369–381, 2019.
https://doi.org/10.1007/978-3-030-15712-8_24

retrieval framework via a weighting function derived from mutual information with respect to a working set of documents. The ranking model can be formulated in terms of query expansion, and thus its implementation is well understood in the broader context of the IR literature.

The work of FZ is worthy of detailed exploration for several reasons: First, axiomatic retrieval is under-explored from a reproducibility perspective, compared to say, BM25 and language modeling approaches. For example, the large-scale study of Lin et al. [4] examined a number of different retrieval models across a number of systems, but did not include any techniques based on axiomatic retrieval. Second, axiomatic semantic term matching provides a strong non-neural baseline, since one of the purported advantages of continuous word representations (on which most neural ranking models depend) is the ability to capture word similarity based on distributional statistics. The importance of FZ has also been recognized by the recent CENTRE reproducibility initiative that cross-cuts CLEF, TREC, and NTCIR. A follow-on paper applying axiomatic semantic term matching to web collections [10] was selected as one of the targets for participants to reproduce. The organizers selected these target papers based on many different factors, including the popularity of the task that the technique tackles, as well as the impact of the work. Although the specific effort we describe here is orthogonal to the CENTRE initiative, the selection of FZ provides independent confirmation that axiomatic semantic term matching represents an important contribution that should be studied in greater detail.

We are able to successfully reproduce the work of FZ using the open-source Anserini information retrieval toolkit built on Lucene. Reproducibility here is used in a precise manner in the sense articulated in recent ACM guidelines,[1] which means "that an independent group can obtain the same result using artifacts which they develop completely independently." Whereas the original FZ paper used Indri, our reimplementation from scratch uses Anserini, sharing no common code. Our implementation, along with detailed documentation and associated run scripts, yields experimental results that are both repeatable (i.e., "a researcher can reliably repeat her own computation") and replicable (i.e., "an independent group can obtain the same result using the author's own artifacts"), both in the sense that ACM defines them (quoted from the ACM guideline referenced above). Given the widespread deployment of Lucene by a large number of organizations in production settings, our implementation increases the options that builders of real-world search applications can explore.

Having reproduced FZ, we conducted additional experiments to generalize the results in several respects: First, we applied the technique to a large number of test collections spanning many different document genres, including newswire, web, and microblogs. Axiomatic semantic term matching is effective for newswire and microblogs, but less so for web collections. Second, we examined a number of parameters that impact effectiveness. In particular, the parameter that determines the weight of semantic matches behaves quite differently across document genres. Also, the technique introduces randomness in the sampling

[1] https://www.acm.org/publications/policies/artifact-review-badging.

of non-relevant documents to construct a working document set–we characterize the impact of this non-determinism. Finally, we demonstrate that although axiomatic semantic term matching was originally developed within the axiomatic retrieval framework, the core ideas can be adapted to other ranking models as well. Specifically, axiomatic semantic term matching also works well on a base ranking model that uses BM25 or query likelihood.

2 Approach

Axiomatic semantic term matching relates document terms that do not match query terms at the lexical level, thus potentially overcoming the vocabulary mismatch problem. In this section, we provide an overview of the technique, borrowing heavily from previous papers [3,10], but refer the reader to those sources for more detailed derivations.

The matching score of term t in a document with respect to query Q comprised of terms $\{q_1, q_2, ..., q_n\}$ is computed as $S(Q,t) = \sum_{q \in Q} s(q,t)/|Q|$, where

$$s(q,t) = \begin{cases} \omega(q) & \text{if } t = q \\ \omega(q) \times \beta \times \frac{\text{MI}(q,t)}{\text{MI}(q,q)} & \text{if } t \neq q \end{cases} \tag{1}$$

For matching terms (i.e., $t = q$), $\omega(q)$ is simply the *idf* of q. In the case of lexical mismatch (i.e., $t \neq q$), the semantic distance between two terms is captured using mutual information (MI) with respect to a working set W (more details below), modulated by β, a parameter that controls how much we "trust" the semantically-related term:

$$\begin{aligned} \text{MI}(q,t) &= I(X_q, X_t|W) \\ &= \sum_{X_q, X_t \in \{0,1\}} p(X_q, X_t|W) \cdot \log \frac{p(X_q, X_t|W)}{p(X_q|W)p(X_t|W)} \end{aligned} \tag{2}$$

Here, X_q and X_t are two binary random variables that denote the presence or absence of term q and term t in the document.

The working set is assembled as follows: First, we take the R top ranked documents from an initial retrieval run, treating them as pseudo-relevant documents. We add to these $(N-1) \times R$ documents (assumed non-relevant) randomly sampled from the collection, excluding the first R documents. This yields a working set comprised of $N \times R$ total documents. Although FZ discuss sampling from external collections, particularly in the web context [10], we do not consider this variation in our study due to limited space.

Considered end to end, the steps involved in axiomatic semantic term matching are as follows:

1. Perform an initial retrieval and construct a working set for computing semantic similarity in the manner described above.

2. For each query term, select the K most similar terms using Eq. (1). From this pool of candidate terms, select the M most similar terms based on $S(Q, t)$.
3. These M terms form the weighted, expanded query. Search the collection with this expanded query and return the final ranked list.

In summary, the parameters for axiomatic semantic term matching are as follows: R, the number of pseudo-relevant documents in the working set; N, which determines the number of additional non-relevant documents to sample, $(N-1) \times R$; K, the cutoff to be considered as a potential expansion term for a query term; M, the total number of expansion terms to add; β, the weight of the expansion terms in Eq. (2).

In our effort, we decided to reproduce axiomatic semantic term matching using Anserini, an open-source information retrieval toolkit built on Lucene [11, 12]. The goal of the Anserini project is to bridge the gap between information retrieval research and real-world search applications, where Lucene has become the *de facto* platform for production deployments. We hope that a Lucene implementation will enable a broader audience (i.e., the open-source community and the long list of companies that run Lucene in production) to try out innovations from academic researchers. The source code of the implementation of axiomatic semantic term matching by Yang and Fang [10] is available online,[2] which provided us with a reference implementation to consult. This implementation is also based on Indri, but it differs from the original implementation in the FZ paper. Due to the availability of this resource, we encountered no difficulties in our implementation efforts.

Beyond reproducing the work of FZ, we explored three research questions to generalize axiomatic semantic term matching:

(RQ1) *Does axiomatic semantic term matching generalize to different types of collections?* The original FZ paper only examined newswire collections, but we experimented with many more test collections spanning three different genres: newswire, web, and microblogs. Many of these collections were not available when the original paper was published.

(RQ2) *How does axiomatic semantic term matching behave with different base ranking models?* Although the formal derivations are couched within the framework of axiomatic retrieval, the operationalization of the model in terms of query expansion means that the technique can be applied to any base ranking model. That is, we can use any number of ranking functions to construct the working set, and use the same ranking function for the expanded query. Our implementation in Anserini makes such explorations easy.

(RQ3) *What is the effect of non-determinism in sampling non-relevant documents?* Semantic term matching weights are computed with respect to a working set populated by sampling (assumed) non-relevant documents from the collection. We examine the impact of this non-determinism on effectiveness.

[2] https://github.com/Peilin-Yang/axiomatic_query_expansion.

3 Experimental Setup

Our experiments used TREC test collections spanning three different genres: newswire, web, and microblogs. The newswire collections are as follows:

- TREC Disks 1 & 2, with topics and relevance judgments from the *ad hoc* task at TREC-1 through TREC-3 (topics 51–200).
- TREC Disks 4 & 5, excluding Congressional Record, with topics and relevance judgments from the *ad hoc* task at TREC-6 through TREC-8 as well as the Robust Tracks from TREC 2003 and 2004.
- The AQUAINT Corpus of English News Text, with topics and relevance judgments from the TREC 2005 Robust Track.
- The New York Times Annotated Corpus, with topics and relevance judgments from the TREC 2017 Common Core Track.

For web collections:

- The WT10g and Gov2 collections from CSIRO (Commonwealth Scientific and Industrial Research Organisation), distributed by the University of Glasgow, with topics and relevance judgments from the web task at TREC-9 for the former, and the Terabyte Tracks at TREC 2004–2006 for the latter.
- The ClueWeb09b and ClueWeb12-B13 web crawls from Carnegie Mellon University, with topics and relevance judgments from the Web Tracks at TREC 2010–2012 for the former and the Web Tracks at TREC 2013 and 2014 for the latter. We did not run experiments on the complete ClueWeb09 and ClueWeb12 collections for two reasons: first, they are too large for running query expansion in practice (i.e., the experiments take too much time), and second, relevance judgments are too sparse to draw firm conclusions (more details later).

And finally, microblog collections:

- The Tweets 2011 collection, with topics and relevance judgments from the TREC 2011 and 2012 Microblog Tracks.
- The Tweets 2013 collection, with topics and relevance judgments from the TREC 2013 and 2014 Microblog Tracks.

All source code for replicating results reported in this paper is available in the Anserini code repository[3] (post v0.3.0 release, based on Lucene 7.6) at commit `08434ad` (dated Jan. 15, 2019).

4 Results

We begin with results from our attempts to directly reproduce the original FZ paper for those collections that overlap with our experimental settings. The original FZ paper, published in SIGIR 2006, predated many of the collections we

[3] http://anserini.io/.

Table 1. Comparisons to the original FZ results (average precision).

	SIGIR 2006		Anserini	
Run	F2EXP	+Ax	F2EXP	+Ax
Robust04	0.2480	0.2850	0.2492	0.2839
Robust05	0.1920	0.2580	0.1985	0.2481

use, and even though FZ report results on other collections, they are generally regarded as either non-standard or too small to support drawing reliable conclusions. Results in terms of average precision are shown in Table 1. Here, F2EXP is used as the base ranking model (the implementation in Anserini, not the Lucene default), with axiomatic semantic term matching denoted by "Ax". In these experiments we used the same parameter settings as in the original paper.

We see that the effectiveness metrics are quite close, despite completely different implementations. The original work of FZ was implemented in Indri, whereas our results are based on Lucene. Differences can be easily be attributed to the document processing pipeline (tokenization, stemming, stopwords, etc.) as well as the inherent non-determinism in constructing the working set (more details below). At a high level, it appears that axiomatic semantic term matching "works as advertised" in terms of effectiveness. Our narrative continues by examining the additional research questions posed in Sect. 2.

As expected, (RQ2) was straightforward to address–our implementation adopts a modular architecture that enabled us to apply different base ranking models for the construction of the working set as well as for the second stage retrieval using the expanded query. In our experiments, in addition to using F2EXP, as the original FZ paper does, we also report results with query likelihood using Dirichlet-smoothed language models (QL) and BM25 (both default Lucene implementations).

In generalizing the results of FZ, our most interesting findings centered around applications to different document genres (RQ1). Furthermore, the parameter of greatest interest is β, which determines the weight of semantically-related terms: we discovered that there are systematic variations across different genres. For these experiments, the remaining parameters were fixed as follows: $N = 30$, $R = 20$, $K = 1000$, $M = 20$. These values represent default settings recommended by FZ. As the original paper already performed a number of parameter explorations, we focused on supplementing those results, since we do not have space for exhaustive examination of all parameters. For these experiments, sampling non-relevant documents was accomplished by setting the random seed to 42, which makes our experiments repeatable.

Results on the newswire collections are shown in Fig. 1: the y axis shows average precision of the top 1000 hits, and the x axis shows the β setting. Each curve denotes a different base ranking model (in a different color): BM25, query likelihood (QL), and F2EXP. The respective baselines without axiomatic semantic term matching are shown as horizontal lines in matching colors. The same plots for the web collections are shown in Fig. 2 (note that we report average

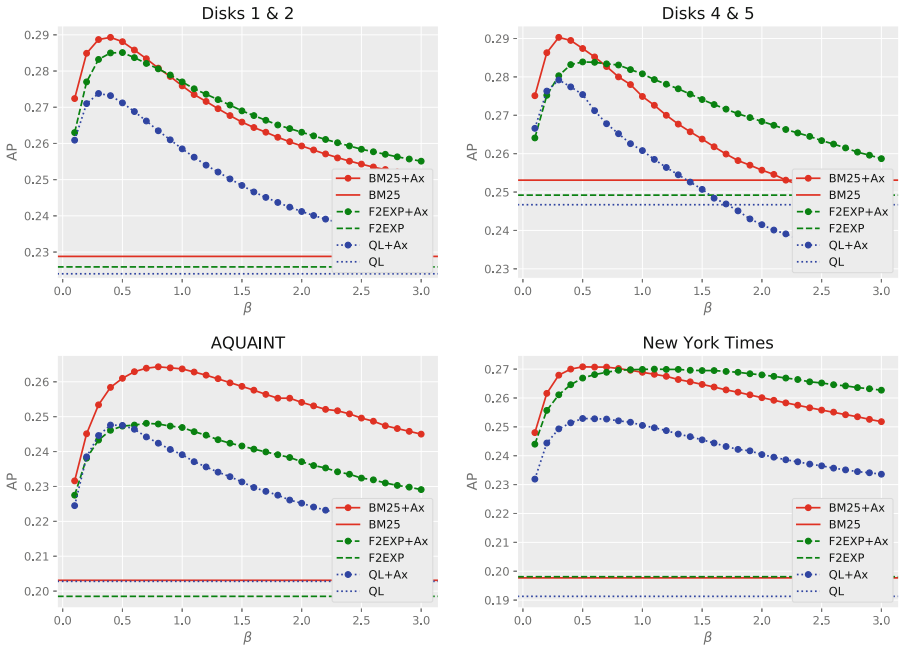

Fig. 1. Results of β tuning experiments on newswire collections.

precision for WT10g and Gov2 but NDCG@20 for the ClueWeb collections, since the shallow pool depths make AP unreliable), and the microblog collections, in Fig. 3. To aid interpretation: $\beta = 1$ places equal importance on both the original query terms and the expansion terms, while $\beta < 1$ means we "trust" expansion terms less (and the opposite for $\beta > 1$).

The newswire collections behave as we would expect—the plots in Fig. 1 are consistent with Fig. 3 in the FZ paper. However, results on the web collections are unexpected: for WT10g and Gov2, axiomatic semantic term matching yields only small improvements in average precision, and only with small values of β. For ClueWeb12-B13, no setting of β improves effectiveness. For the microblog collections, we also observe qualitatively different behavior: First, optimal effectiveness is reached at a larger value of β, which means that the ranking model places more importance on expansion terms. Second, effectiveness does not appear to be very sensitive to β at all. Whereas average precision decays sharply with larger values of β on newswire collections, effectiveness decays much more slowly for microblogs.

Before drawing any firm conclusions from these results, we need to rule out evaluation artifacts. One obvious culprit is unjudged documents—query expansion has the possibility of retrieving documents that are not part of the original evaluation pool. Figure 4 shows the results of this analysis for BM25. For each collection, we plot the fraction of unjudged documents in the top 20, 50, and 100 hits. The first row shows results for the newswire collections, the second row for

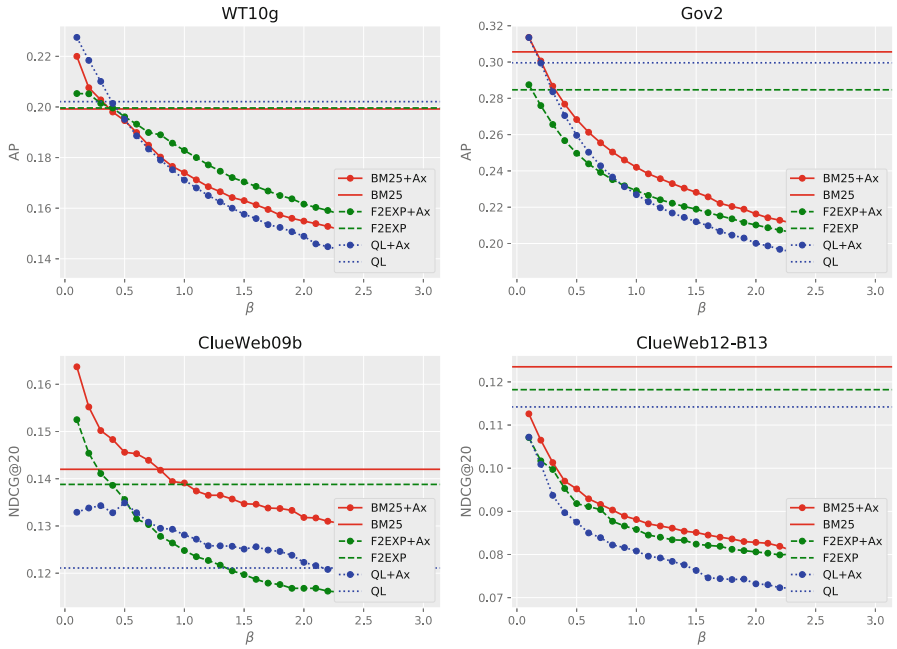

Fig. 2. Results of β tuning experiments on web collections.

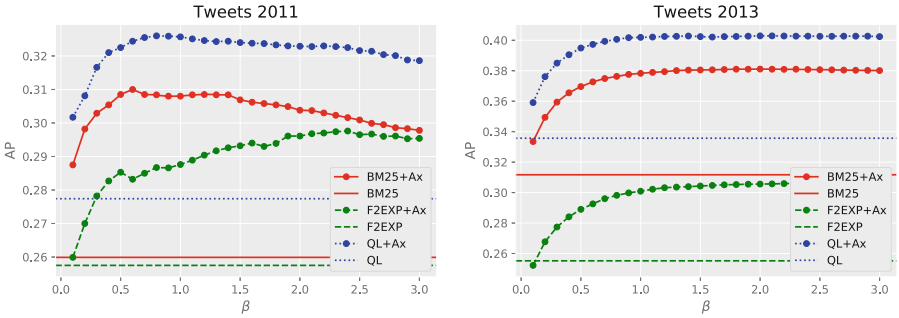

Fig. 3. Results of β tuning experiments on microblog collections.

the web collections,[4] and the third row for the microblog collections. Ideally, the fraction of unjudged documents should be constant across different β settings; that is, no setting should be penalized by retrieving more unjudged documents. The absolute value of missing judgments is less important, since judgments will always be incomplete in any pooling-based test collection. Instead, we are more interested in whether different settings of β are unfairly penalized.

[4] For the ClueWeb collections, we measured effectiveness in terms of NDCG@20, so the analysis for the top 50 and 100 documents are not applicable; nevertheless, we have included those results in the graphs for completeness.

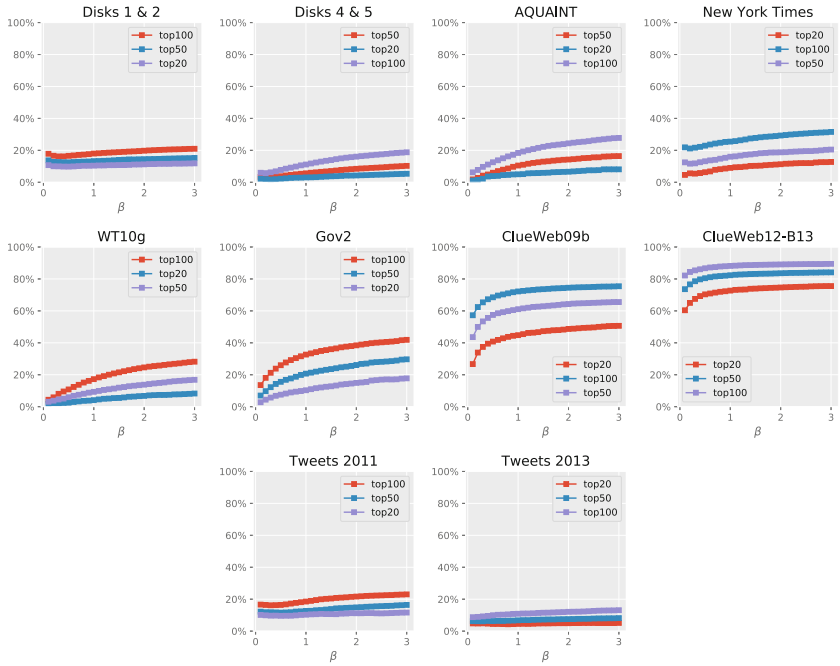

Fig. 4. Analysis of missing judgments using BM25 as the base ranking model.

Results from Disks 1 & 2 are closest to our ideal—the fraction of missing judgments does not vary much across β settings (and furthermore, the absolute values are quite low). For the newswire collections, the results on AQUAINT (Robust05) deviate the most from our ideal—for example, a setting of $\beta = 1$ yields around 10% *more* unjudged documents vs. $\beta = 0.5$ at rank 100. For web collections (second row in Fig. 4), we observe even more missing judgments. For ClueWeb12-B13, with any setting of β, over 60% of the documents are unjudged. The microblog test collections are reasonably well behaved, where the fraction of missing judgments is comparable to newswire collections.

Given the evidence presented above, the following conclusions are supported with respect to (RQ1): axiomatic semantic term matching appears to be effective across a range of newswire collections with $\beta = 0.5$; the technique also appears to be effective for microblog collections, with a setting of $\beta = 1.0$. These β values should be taken as rough, coarse-grained guides. In fact, we argue that fine-grained tuning is essentially meaningless due to missing judgments and the fact that effectiveness differences are not very large in a broad range around the above-proposed settings. For web collections, a setting of $\beta = 0.1$ yields slightly better effectiveness in some cases, but however, there is insufficient evidence to decide between two competing hypotheses: That axiomatic semantic term matching is not effective for web collections, or that current evaluation resources are unable to accurately determine its effectiveness. If the former turns out to be true, *why* would be an interesting follow-on question.

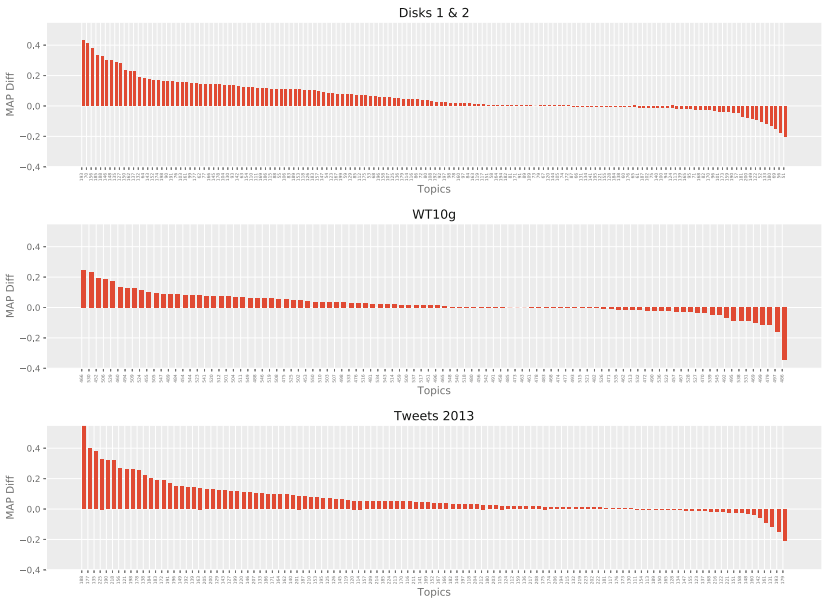

Fig. 5. Per-topic analysis for Disks 1 & 2 (top), WT10g (middle), and Tweets 2013 (bottom) comparing axiomatic semantic term matching with baseline BM25 ranking.

We attempted to dig a bit deeper into understanding the behavior of axiomatic semantic term matching across different document genres by analyzing per-topic effectiveness differences. Figure 5 shows results for a representative collection from each genre: Disks 1 & 2, WT10g, and Tweets 2013. These collections were selected because they contained the fewest unjudged documents according to the analysis in Fig. 4, thus affording us the greatest confidence in the effectiveness measurements. Each bar represents a topic and its height captures the average precision difference between baseline BM25 and axiomatic semantic term matching with BM25 as the base ranking model. Bars are sorted in descending order of effectiveness differences, from left to right, where negative bars represent topics where axiomatic semantic term matching hurts effectiveness.

As is typical of many query expansion techniques, axiomatic semantic term matching helps some topics but hurts other topics. The relative proportion of the beneficial vs. detrimental cases does not seem markedly different across genres, but it appears that even for the best topics in WT10g, the technique does not help as much as in the other two collections. Also, for WT10g, the worst-performing topics have decreases in AP that are greater than in the other collections. We followed up with manual analysis of the worst-performing topics across all three collections, comparing the original queries with the expanded queries. Unfortunately, this did not reveal any obvious insights. For example, we hypothesized that since web collections contain more noisy text, the quality of the expansion terms might be worse. However, this was not the case—the expansion terms all appeared reasonable and their quality was not markedly different from query expansion terms in the other two collections.

In answering (RQ2), looking across newswire, web, and microblog collections, it seems clear that axiomatic semantic term matching can be applied to a variety of base ranking models. For the newswire collections, effectiveness appears to be highest using BM25, with F2EXP slightly better than QL in most cases. For the web collections, the effectiveness of all three ranking models is quite similar. For the microblog collections, we observe large differences in average precision, but these results are consistent with known characteristics of the collections: BM25 does not work well for ranking microblogs because posts do not differ much in length, and thus the length normalization factor in the scoring function has little impact. For the TREC Microblog Tracks, QL is the preferred baseline [5].

Our final set of experiments tackled (RQ3) and examined the inherent non-determinism involved in the construction of the working set when sampling non-relevant documents. These experiments used the β recommendations above with the same settings of the other parameters. For each test collection, using the BM25 base ranking model, we repeated the ranking experiments 100 times with different random seeds.[5] The results are summarized in box-and-whiskers plots in Fig. 6, which report average precision except for the ClueWeb collections, which show NDCG@20. The blue dotted line in each case represents the effectiveness

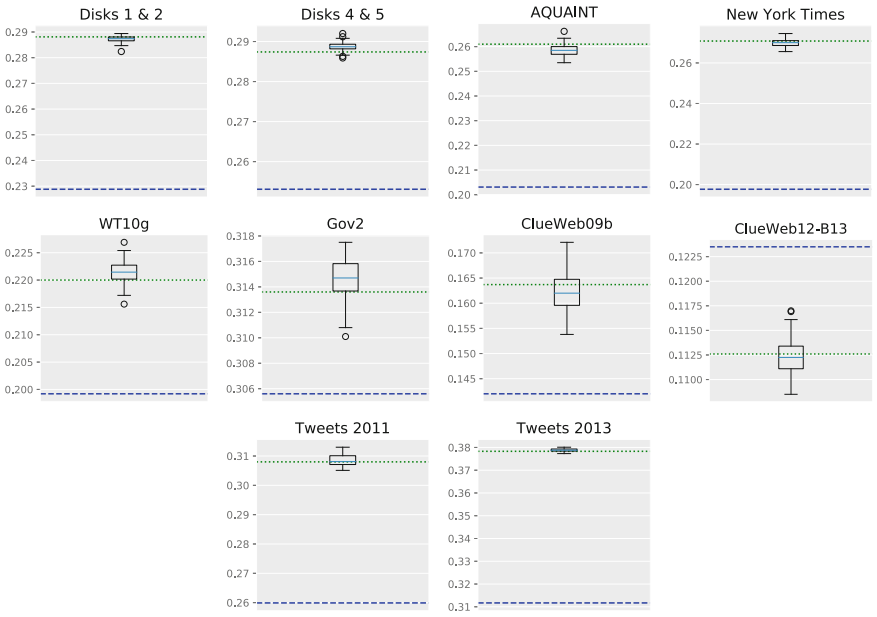

Fig. 6. Box-and-whiskers plots showing the distribution of scores across 100 random seeds when sampling non-relevant documents to construct the working set, with BM25 as the base ranking model. The BM25 baselines are shown as blue dotted lines, while the single-point measurements are shown as green dotted lines.

[5] This was accomplished by using 42 as the "meta seed" to generate a pseudo-random sequence of random seeds for each experimental run.

of the BM25 baseline, and the green dotted line represents the single-point effectiveness measurement from the comparable experiments above.

To specifically answer (RQ3): We observe that the variations in effectiveness that can be attributed to random seed selection is quite small, and that even the low effectiveness outliers are well above the BM25 baselines for both newswire and microblog collections. For both document genres, the single-point effectiveness measurement is within the range predicted by the box-and-whiskers distributions. We can conclude that axiomatic semantic term matching is robust with respect to document sampling for the working set. The results for the web collections are consistent with the findings above, and suggest that axiomatic semantic term matching helps for three of the four collections. For ClueWeb12-B13, the large fraction of unjudged documents prevents us from drawing any meaningful conclusions, as discussed above.

5 Conclusions

We have successfully reproduced the axiomatic semantic term matching work of Fang and Zhai in Anserini, based on the popular open-source Lucene search engine. The work is over a decade old, and this paper generalizes the techniques to web and microblog collections, beyond the newswire collections in the original paper. We confirm that axiomatic semantic term matching is indeed effective on newswire, and that microblogs similarly benefit. However, the effectiveness of these techniques on web collections is unclear; we are unable to draw any firm conclusions due to limitations of existing test collections (too many unjudged documents). Nevertheless, it is clear that different document genres require different weights on the importance of semantic term matches, although there does not appear to be any principled rationale for those settings.

All of the code necessary to replicate the experiments reported in this paper is available in the Anserini open-source IR toolkit. Already contributed to our code repository are numerous models frequently used in academic information retrieval research, including relevance models and sequential dependence models. Our longer term hope is that Lucene-based implementations bring academia and industry into better alignment, allowing researchers an easier path to achieve real-world impact via deployments of real-world search applications.

Acknowledgments. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

1. Berger, A., Lafferty, J.: Information retrieval as statistical translation. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 222–229. SIGIR 1999. ACM, New York (1999). https://doi.org/10.1145/312624.312681

2. Fang, H., Zhai, C.: An exploration of axiomatic approaches to information retrieval. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 480–487. SIGIR 2005. ACM, New York (2005). https://doi.org/10.1145/1076034.1076116
3. Fang, H., Zhai, C.: Semantic term matching in axiomatic approaches to information retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 115–122. SIGIR 2006. ACM, New York (2006). https://doi.org/10.1145/1148170.1148193
4. Lin, J., et al.: Toward reproducible baselines: the open-source IR reproducibility challenge. In: Ferro, N., et al. (eds.) ECIR 2016. LNCS, vol. 9626, pp. 408–420. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30671-1_30
5. Lin, J., Efron, M.: Overview of the TREC-2013 Microblog Track. In: Proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013), Gaithersburg, Maryland (2013)
6. Onal, K.D., et al.: Neural information retrieval: at the end of the early years. Inf. Retrieval **21**(2–3), 111–182 (2018). https://doi.org/10.1007/s10791-017-9321-y
7. Rocchio, J.J.: Relevance feedback in information retrieval. In: Salton, G. (ed.) The SMART Retrieval System-Experiments in Automatic Document Processing, pp. 313–323. Prentice-Hall, Englewood Cliffs (1971)
8. Voorhees, E.M.: Query expansion using lexical-semantic relations. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 61–69. SIGIR 1994. ACM, New York (1994). http://dl.acm.org/citation.cfm?id=188490.188508
9. Xu, J., Croft, W.B.: Improving the effectiveness of information retrieval with local context analysis. ACM Trans. Inf. Syst. **18**(1), 79–112 (2000)
10. Yang, P., Fang, H.: Evaluating the effectiveness of axiomatic approaches in web track. In: Proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013), Gaithersburg, Maryland (2013)
11. Yang, P., Fang, H., Lin, J.: Anserini: enabling the use of Lucene for information retrieval research. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1253–1256. SIGIR 2017. ACM, New York (2017). https://doi.org/10.1145/3077136.3080721
12. Yang, P., Fang, H., Lin, J.: Anserini: reproducible ranking baselines using Lucene. J. Data Inf. Qual. **10**(4) (2018). Article 16