Exploring Diversification in Non-factoid Question Answering

Lakshmi Vikraman, W. Bruce Croft and Brendan O'Connor
University of Massachusetts Amherst, Amherst, MA, USA
{lvnair,croft,brenocon}@cs.umass.edu

ABSTRACT

Retrieving short, precise answers to non-factoid queries is an increasingly important task, especially for mobile and voice search. Many of these questions may have multiple or alternative answers. In an environment where answers are presented incrementally, this raises the question of how to generate a diverse ranking to cover these alternatives. Existing search diversification algorithms generate diverse document rankings using explicit or implicit methods based on topical similarity. The goal of this paper is to evaluate the impact of applying these existing document diversification frameworks to the problem of answer diversification to determine if topical diversity is related to answer diversity. Using two common diversification algorithms, xQUAD and PM-2, and three question answering test collections, we show that topic diversification can help to generate more effective rankings but is not consistent across different queries and test collections.

ACM Reference Format:
Lakshmi Vikraman, W. Bruce Croft and Brendan O'Connor. 2018. Exploring Diversification in Non-factoid Question Answering. In *ICTIR '18: 2018 ACM SIGIR International Conference on the Theory of Information Retrieval, September 14–17, 2018, Tianjin, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3234944.3234973

1 INTRODUCTION

Current search engines provide answers to queries in the form of SERPs (Search Engine Result Pages), which are a series of web page links ranked by their relevance to the query. While this is effective, it compels the user to scroll through the links and identify pertinent answers to the query. It is especially inconvenient in scenarios such as mobile search where the display space is limited. Recent retrieval models in IR address this issue by focusing on short answers instead of documents. Some queries can be answered by short entity level answers while the majority of queries are more open ended with answers spanning multiple sentences. The former class of queries is factoid queries while the latter is called non-factoid queries. Examples of the two types of queries are shown below.

- **Factoid query:** What is the current temperature in New York?
- **Non-factoid query:** How is the hurricane season in the Houston area?

Non-factoid queries may have multiple answers of the same or different types associated with them. An example of the former case is the query "`Examples of anthrax hoax`", where the user might view the first instance of the "hoax" and then wish to view other alternative examples. In the latter case where the answers are of different types, for example, "`What is the best diet to prevent cancer`", the user might be interested in all the relevant answers. For these types of queries, the diversified lists are useful to collate and organize the answers.

Significant research has been done in the area of Search Result Diversification of documents. This has also been covered under TREC Web Track's diversity tasks ([3, 16]). An example TREC query (topic 111) from TREC Web Track 2011 is shown in Table 1.

Table 1: TREC query

TREC query	lymphoma in dogs
Subtopic 1	What treatments are available for dogs diagnosed with lymphoma?
Subtopic 2	What are the symptoms of lymphoma in dogs?
Subtopic 3	What are the risk factors or causes of lymphoma in dogs?

For these types of queries, topical diversification has been found to be effective in producing a list of documents covering the various subtopics. An example is the query in Table 1, where the diversified document list would cover "causes", "symptoms" and "treatments" of lymphoma. However, our focus is on more specific queries with multiple possible answers, which potentially cover only a single subtopic. For instance in the query "`What are the symptoms of lymphoma in dogs?`", the subtopic is "symptoms" and the diversified answer list should cover all symptoms of the disease. The question that we address in this paper is whether techniques used for diversifying documents would be effective in producing a diversified list of answers for non-factoid queries.

Standard diversification algorithms are categorized into two types: implicit and explicit. The first type assumes that each document represents its own topic and diversifies based on document similarity. These classes of methods do not attempt to cover underlying query aspects explicitly, which make them less effective in practice. An example of the implicit approach is Maximal Marginal Relevance (MMR) [1]. On the other hand, the explicit approach models the query topics and diversifies the ranked list based on topic coverage. The topics are generated automatically from the initial retrieved set [7] or created manually. Examples include xQUAD [15] and PM-2 [8]. We select a topical diversification framework: term level diversification [7] that uses topic terms to perform diversification.

To determine if topical diversification can be directly used to create a diversified answer list, we apply it to three existing non-factoid question answering collections: WebAP [10], nfl6 [4] and

ICTIR '18, September 14–17, 2018, Tianjin, China

ACM ISBN 978-1-4503-5656-5/18/09…$15.00
https://doi.org/10.1145/3234944.3234973

WikiPassageQA [6]. WebAP contains multiple answers per query and the terms generated by the algorithm can be used to create the diversified answer list. However, the other collections (nfl6 and WikiPassageQA) contain only one judged relevant answer per query, and diversified lists cannot be evaluated effectively. Hence, for such cases we focus on determining if diversity can improve the ranking of the relevant answer in the ranked list. Due to the lack of manually annotated diversified answers for these datasets, we evaluate them based on standard relevance metrics instead of using diversification metrics.

2 RELATED WORK

Non-factoid Question Answering: Research in the area of non-factoid question answering is still relatively new. Recent work using deep learning techniques has shown to be more effective for this task. Wang and Nyberg [17] used a BiLSTM model with word2vec [14] word embeddings to retrieve answers to non-factoid questions. Cohen and Croft [4] demonstrated the impact of updating word embeddings during the training process to improve performance. More recently Cohen and Croft [5] used character level embeddings for this task.

Implicit Diversification: Implicit diversification assumes that the diverse information needs can be satisfied by dissimilar documents. One of the most popular implicit approaches is MMR [1]. More recently various supervised implicit techniques have been proposed. Zhu et al. [19] proposed a model to optimize both novelty and relevance. Xia et al. [18] proposed a neural tensor model which learns document representations automatically and does not require handcrafted features.

Explicit Diversification: Explicit approaches model query subtopics explicitly and generate a ranked list, which is optimized based on the coverage of these topics. xQUAD and PM-2 are examples of this approach where xQUAD [15] uses query reformulations as topics and PM-2 [8] is a proportionality based diversification model. Hu et al. [9] proposed a hierarchical variant of xQUAD and PM-2 which models topics as a hierarchy instead of a list.

3 TERM LEVEL DIVERSIFICATION

In this paper, we use term level diversification introduced by Van and Croft [7]. We first introduce some notation. Let q be a query, $T = \{t_1, t_2, ...t_n\}$ be the topically diverse set of terms corresponding to the query , $W = \{w_1, w_2, ...w_n\}$ be the weights for the terms, $R = \{a_1, a_2, ...a_m\}$ be the initial ranked list of answers and $S = \{a_1, a_2...a_m\}$ be the diversified ranked list. The aim is to generate the set S using a diversification framework given inputs q,T, W, R. Since we don't have a predefined set of topic terms T per query, we use the algorithm *DSPApprox* proposed by Lawrie and Croft [11, 12] to identify them automatically.

3.1 DSPApprox Algorithm

We first identify a set of vocabulary terms $V = v_1, v_2, ...v_k$ from the answers in the initial ranked list R. The terms are categorized into three types - unigrams, unigrams tagged as nouns and noun phrases. To reduce noise, we discard all terms (1) which do not occur in at least two answers (2) have less than two characters (3) numbers. From the vocabulary, all terms that occur within a window of size w of the query terms are selected as candidate topic terms T. A score is calculated for each of these terms based on two measures: topicality and predictiveness. Topicality of a term measures how well it describes a set of documents.

$$TP(t) = P_R(t|q) log_2 \frac{P_R(t|q)}{P_c(t)} \tag{1}$$

where $P_R(t|q)$ is the relevance model estimated from R and $P_c(t)$ is the language model for the entire data collection. Predictiveness of a term(t) measures how well the term predicts the surrounding neighbors v within a window of size c and Z is the normalization factor which is set as the size of the vocabulary.

$$P_R(t) = \frac{1}{Z}\sum_{v} P(t|v) \tag{2}$$

The score of each candidate term is the product of topicality and predictiveness: $P_R(t)TP(t)$. At each step of the algorithm, the highest scoring term is added to the output set O and the predictiveness score of other candidate terms, which predicts the same vocabulary, is reduced.

3.2 Diversification Algorithms

xQUAD: The first diversification framework we chose is xQUAD [15]. The general framework for xQUAD is given below.

$$a^* = \underset{a_j \in R}{\text{argmax}}(1-\lambda)Rel(a_j, q) + \lambda D(a_j, S) \tag{3}$$

The xQUAD score is a linear combination of relevance and diversity scores where the relevance score is the query likelihood score and diversity score is based on how well the answers cover the topic terms which have not already been covered by set S.

PM-2: The second diversification algorithm we experimented with is the proportionality based model PM-2 [8]. The answers are assigned to topics to best maintain proportionality of the ranked list. Initially we calculate the quotient qt_i for each topic term t_i.

$$qt_i = \frac{w_i}{2s_i + 1} \tag{4}$$

where s_i is the portion of answers assigned to each topic term t_i. The topic term t_{i*} with the largest qt_i is selected as the topic term to be covered and the answer to cover this topic is selected as follows.

$$a^* = \underset{a_j \in R}{\text{argmax}}\, \lambda qt_{i*} P(a_j|t_{i*}) + (1-\lambda) \sum_{i \neq i*} qt_i P(a_j|t_i) \tag{5}$$

After selecting a^*, the portion of answers assigned to each topic term t_i is increased and the answer a^* is placed in set S. The ranking of each answer depends on the order in which it is placed.

4 EXPERIMENTAL SETUP

Data Overview: Three datasets are used for the experiments namely nfl6, WebAP and WikiPassageQA.

- **nfl6:** nfl6 introduced in Cohen et al. [4] is a subset of Yahoo's Webscope L6 collection consisting of 87362 questions and their corresponding answers. This dataset was created by excluding potential factoid questions from the L6 collection. Each question has a best answer associated with it along with a set of additional answers submitted by the users.

- **WebAP:** WebAP dataset was introduced by Keikha et al. [10]and consists of 82 queries. The corresponding answers to these queries were created from the top 50 retrieved documents in the GOV2 collection. Each document is split into relevant and non-relevant sections with the non-relevant section further subdivided into random sized non-overlapping passages. The relevant passages with relevance judgments of "Perfect","Excellent","Fair" and "Good" are mapped to 4~1 respectively and the non-relevant passages mapped to 0.
- **WikiPassageQA:** WikiPassageQA dataset introduced by Cohen et al. [6] consists of 4162 queries created from 863 Wikipedia articles. Here, each query corresponds to a single article though an article can have multiple queries associated with it. Each article was split into passages consisting of six sentences each. Hence, the relevant answer might be split among multiple passages.

Baseline Retrieval Model: The initial retrieval run is obtained using the query likelihood model implemented via the Indri search engine ($\mu = 2500$). This provides a set of ranked answers, which is the input to the diversification systems. For nfl6, since the set of answers for a query is very low, the retrieval is done over the entire collection. A query is discarded if the relevant answer is not present in the top 100 retrieved set which results in a final subset of 31229 queries. In case of WebAP and WikiPassageQA, the initial retrieval run is over the passages within the documents annotated for a particular query.

Model settings : For the various models, a prior topic weight w_i is calculated. In the term diversification model this is set to be the utility score given by DSPApprox algorithm. In LDA topic based models, the weights are assumed to be uniform. In the term level diversification model, we expand the topic terms with query terms while calculating $P(a|t)$ to reduce noise. To investigate the impact of hyperparameters, we performed grid search over the entire dataset and found the variation in performance insignificant.

Baselines:

- **Query Likelihood:** The first baseline is the Query Likelihood Model with default Dirichlet prior smoothing ($\mu = 2500$).
- **MMR Diversification:** The second baseline is MMR(Maximal Marginal Relevance) [1]. This is an implicit diversification model, which performs diversification based on document similarity.

$$a^* = \underset{a_j \in R}{\operatorname{argmax}}(1-\lambda)Rel(a_j, q) + \lambda \max_{a_i \in S} sim(a_j, a_i) \qquad (6)$$

  The similarity function used is cosine similarity and the answers are represented as sparse vectors where each dimensions corresponds to the term frequency within the answer.
- **LDA based Diversification:** The third baseline employs a topic modeling approach to perform diversification. This model was first proposed by Carterette and Chander [2] and uses LDA to generate topics. We use the mallet implementation [13] of LDA, which also outputs a document-topic score that is used directly in the diversification framework.

Evaluation Metric: We use standard relevance metrics such as Precision, NDCG (Normalized Discounted Cumulative Gain) and MRR (Mean Reciprocal Rank) instead of diversification metrics due to lack of annotated data. Statistical significance is measured using the paired two-tailed t-test with p-value < 0.05.

Table 2: Example topic terms for queries

Query	Topic Terms
What does obsessive compulsive behavior mean?	disorder,compulsive,time, ocd,thought
How do nutrients get transferred to muscle?	food,cell,weight,diet,body

5 RESULTS AND ANALYSIS

Table 3 reports results for the various models on the three datasets. For nfl6 and WikiQA, term level diversification approach performs better than other diversification baselines (MMR and LDA), but performs slightly worse than the QL model. WebAP queries improved over diversification baselines and also showed a small improvement over QL, however the improvements were not statistically significant. Between xQUAD and PM-2 based methods, xQUAD performed much better since it scores answers based on relevance and diversity while PM-2 scores are only based on diversity. In general, the explicit topic based methods outperformed the implicit MMR model.

Impact of keyterms on answer diversity: The diversification algorithms generate answer rankings based on the terms extracted by the DSPApprox algorithm. In general, it was observed that the algorithm improves the rankings of the answers, which contain these keyterms. Hence the correctness of the terms is crucial for the success of this method. The topic terms for the first query in the Table 2 are related to "`OCD`" which helps in retrieving the relevant answers. The topics terms for the second query, while highly related to "`body`" do not contain the term "`bloodstream`" which is crucial to finding the correct answer.

Collection specific observations:

- **WebAP:** WebAP consists of queries with multiple answers associated with them. Overall, the term level diversification framework seems to work better on this dataset than others. An example of a WebAP query is "`method control type ii diabetes`" whose topic terms are "`diabete,patient,care,study,al`".

After calculating Win Loss statistics with respect to NDCG@10 on the *DSP Unigram xQUAD* model, it was found that out of 80 queries, 53 queries have the same NDCG value as the QL baseline with improvements for 14 queries and a decrease in value for 13 queries. On analyzing this further, it was found that the topic terms were not fine-grained enough to cover the various answers in many cases. The specific methods to control "diabetes" are not reflected in the topic terms which in turn return a re-ranked answer list consisting of these generic terms associated with the question. However, despite the lack of specificity, the generic terms help in improving some of the queries.

- **WikiPassageQA:** WikiPassageQA dataset consists of queries with a single relevant answer passage where multiple queries could be associated with the same wikipedia page. The keyterms are extracted from the page associated with the query. The overlap of the text in the initial ranked lists as well as the high topical similarity between the queries contribute to identification of topic terms which do not effectively discriminate between them. For example, the two queries "`How does the Malaysian political system account for the multiethnic nature of the country?`" and "`How do Malaysia combine English common law and Sharia law?`"

Table 3: Results on WebAP, nfl6 and WikiPassageQA datasets. The models prefixed with "DSP" refer to the various variants of the term diversification model [7].

Model	WebAP			nfl6			WikiPassageQA		
	NDCG@10	P@10	MRR	NDCG@10	P@10	MRR	NDCG@10	P@10	MRR
QL	0.1122	0.1262	0.2404	0.3546	0.0556	0.3085	0.5766	0.1136	0.5947
MMR	0.1061	0.1263	0.2200	0.2368	0.0315	0.2304	0.4325	0.0731	0.5203
LDA	0.0809	0.0975	0.1930	0.2987	0.0506	0.2576	0.5174	0.1074	0.5478
DSP Unigram xQUAD	0.1167	0.1313	0.2354	0.3499	0.0547	0.3053	0.5724	0.1126	0.5938
DSP Noun xQUAD	0.1167	0.1313	0.2354	0.3534	0.0554	0.3075	0.5720	0.1125	0.5937
DSP Phrase xQUAD	0.1137	0.1275	0.2374	0.3487	0.0544	0.3047	0.5627	0.1127	0.5880
DSP Unigram PM-2	0.1027	0.1175	0.2107	0.3146	0.0501	0.2747	0.5270	0.1071	0.5394
DSP Noun PM-2	0.1026	0.1175	0.2107	0.3141	0.0500	0.2743	0.5268	0.1069	0.5398
DSP Phrase PM-2	0.0855	0.1037	0.1680	0.2645	0.0448	0.2280	0.4955	0.1040	0.4999

generated from the wikipedia page "Malaysia" have the same set of topic terms "`malay,malaysian,state,peninsular,government`" which could be misleading.

- **nfl6:** nfl6 dataset has a single relevant answer per query.For such collections, it is expected that the topic terms would help differentiate between the relevant and the non-relevant answers. However this dataset has a couple of issues, which makes this hard. Many of the questions are very open-ended lacking specificity in answers. For example, `"What is the difference between insanity and genius?"` is an ambiguous question, which is hard even for humans to answer. Another issue arises from the way the data was collected. Since the data was created automatically from a public forum, there are cases where multiple answers could be deemed relevant by the model but are not marked as such.For example, for the query "`How to cure arthritis?`", the relevant answer is "`I'm not sure there is a cure. I've had arthritis for a long time. They told me to lose weight, which I did and it felt worse. My grandpa swears by cod liver oil though, try it and see if it helps`". While there might be multiple ways to cure "arthritis", only one answer is marked as relevant, and even if the model assigns a high score for these alternative answers, they are not considered "relevant" during evaluation.

6 CONCLUSION AND FUTURE WORK

This paper explores the impact of applying a term level diversification model to the task of answer diversification. As observed, this technique helps for queries where the topic terms overlap with the terms in the relevant answers. Overall, for the dataset with multiple relevant answers (WebAP), this method helps, though the improvements are not statistically significant. In case of datasets with single relevant answer, it does not outperform the QL baseline. However, the term level diversification method performs better than other diversification baselines such as MMR and LDA diversification models. During analysis, it was observed that for many of the queries in these datasets, the differences between the relevant and non-relevant answers were too subtle to be captured by the algorithm. In future, we plan to explore deep learning models to capture these subtleties due to its ability to model long term dependencies and semantic relationships between terms.

ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-1715095. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor. The authors thank Myungha Jang and John Foley for their valuable comments and Daniel Cohen for providing the datasets.

REFERENCES

[1] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*. ACM, 335–336.

[2] Ben Carterette and Praveen Chandar. 2009. Probabilistic models of ranking novel documents for faceted topic retrieval. In *CIKM*. ACM, 1287–1296.

[3] Charles L Clarke, Nick Craswell, and Ian Soboroff. 2009. *Overview of the trec 2009 web track.* Technical Report. WATERLOO UNIV (ONTARIO).

[4] Daniel Cohen and W Bruce Croft. 2016. End to end long short term memory networks for non-factoid question answering. In *ICTIR*. ACM, 143–146.

[5] Daniel Cohen and W Bruce Croft. 2018. A Hybrid Embedding Approach to Noisy Answer Passage Retrieval. In *ECIR*. Springer, 127–140.

[6] Daniel Cohen, W. Bruce Croft, and Liu Yang. 2017. WikiPassageQA: A Benchmark Collection for Research on Non-factoid Answer Passage Retrieval.. In *SIGIR*.

[7] Van Dang and Bruce W Croft. 2013. Term level search result diversification. In *SIGIR*. ACM, 603–612.

[8] Van Dang and W Bruce Croft. 2012. Diversity by proportionality: an election-based approach to search result diversification. In *SIGIR*. ACM, 65–74.

[9] Sha Hu, Zhicheng Dou, Xiaojie Wang, Tetsuya Sakai, and Ji-Rong Wen. 2015. Search result diversification based on hierarchical intents. In *CIKM*. ACM, 63–72.

[10] Mostafa Keikha, Jae Hyun Park, and W Bruce Croft. 2014. Evaluating answer passages using summarization measures. In *SIGIR*. ACM, 963–966.

[11] Dawn Lawrie, W Bruce Croft, and Arnold Rosenberg. 2001. Finding topic words for hierarchical summarization. In *SIGIR*. ACM, 349–357.

[12] Dawn J Lawrie and W Bruce Croft. 2003. Generating hierarchical summaries for web searches. In *SIGIR*. ACM, 457–458.

[13] Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. (2002). http://mallet.cs.umass.edu.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*. 3111–3119.

[15] Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting query reformulations for web search result diversification. In *WWW*. ACM, 881–890.

[16] Ian M Soboroff, Nick Craswell, Charles L Clarke, and Gordon Cormack. 2011. *Overview of the trec 2011 web track.* Technical Report.

[17] Di Wang and Eric Nyberg. 2015. A long short-term memory model for answer sentence selection in question answering. In *ACL*, Vol. 2. 707–712.

[18] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2015. Learning maximal marginal relevance model via directly optimizing diversity evaluation measures. In *SIGIR*. ACM, 113–122.

[19] Yadong Zhu, Yanyan Lan, Jiafeng Guo, Xueqi Cheng, and Shuzi Niu. 2014. Learning for search result diversification. In *SIGIR*. ACM, 293–302.