Result Diversification for Tweet Search

Makbule Gulcin Ozsoy, Kezban Dilek Onal, and Ismail Sengor Altingovde

Middle East Technical University, Ankara, Turkey
{makbule.ozsoy,dilek,altingovde}@ceng.metu.edu.tr

Abstract. Being one of the most popular microblogging platforms, Twitter handles more than two billion queries per day. Given the users' desire for fresh and novel content but their reluctance to submit long and descriptive queries, there is an inevitable need for generating diversified search results to cover different aspects of a query topic. In this paper, we address diversification of results in tweet search by adopting several methods from the text summarization and web search domains. We provide an exhaustive evaluation of all the methods using a standard dataset specifically tailored for this purpose. Our findings reveal that implicit diversification methods are more promising in the current setup, whereas explicit methods need to be augmented with a better representation of query sub-topics.

Keywords: Microblogging, Tweet search, novelty, diversity.

1 Introduction

Microblogging sites have recently become world-wide popular platforms for sharing and following emerging news and trending events, as well as expressing personal opinions and feelings on a wide range of topics. In Twitter, a prominent example of such platforms, hundreds of millions of users post 500 million tweets per day. In addition to reading tweets in their own feed, Twitter users often conduct search on the posted content. As of 2014, the number of queries submitted to Twitter per day is reported to be more than two billion [2].

The nature of search in Twitter is different from that of the typical Web search in several ways. Twitter users are more interested in searching other people (especially celebrities) or trending events (usually expressed via hashtags) and more likely to repeat the same query to monitor the changes in the content in time [23]. Given the users' desire on the timely and novel content, earlier works essentially focus on filtering near-duplicates in search results, which might be abundant due to retweeting or posting of the same/similar content by several users in the same time period. A complementary issue, which is mostly overlooked in the literature, is the diversification of the search results, i.e., covering different aspects (sub-topics) of the query topic in the top-ranked results.

In typical web search, diversification of results is usually needed due to the ambiguous or vague specification of queries. In case of Twitter, queries are even shorter (1.64 words on the average [23]), which indicates a similar need for diversification. Furthermore, due to the bursty nature of the microblogging, some

B. Benatallah et al. (Eds.): WISE 2014, Part II, LNCS 8787, pp. 78–89, 2014.

content regarding a particular query aspect can be quickly buried deep in the tweet stream, if the user is not fast enough to see it. For instance, assume a query "WISE2014", with possible different aspects that may discuss the technical coverage of the conference (accepted papers, etc.), logistics (visa issues, travel arrangements, hotels, etc.) and social events during the conference. At the time a user submits this query, it might be possible that the other users are essentially posting about, say, the logistics issues, which are not interest of the searcher if she does not plan to attend. In this case, if she has no patience or time to scroll down tens of tweets, she would miss the earlier messages abut the accepted papers that she is really interested in. Note that, diversification of search results does not only help the end-users to quickly grasp different dimensions of the topic in question, but it may also improve applications that submit queries to the Twitter API and retrieve a few top-ranked results for further processing.

In this paper, we address the result diversification problem for tweet search. To this end, we adopt various methods from the literature that are introduced for text summarization and web search result diversification. In particular, we consider LexRank [8] and Biased LexRank [14] as representative methods from the field of text summarization, and several implicit and explicit diversification methods, namely, MMR [3], Max-Sum [10], Max-Min [10] and xQuAD [19] from web search domain. For comparison, we also include the Sy method proposed in the context of near-duplicate elimination in tweet streams [21,22]. While investigating the performance of these approaches, we analyze the impact of various features in computing the query-tweet relevance and tweet-tweet similarity scores. We evaluate the performance using a benchmark collection, the Tweets2013 corpus [22], recently released for this purpose. To the best of our knowledge, apart from the Sy method [22], none of these methods were employed in the context of tweet search diversification and evaluated in a framework involving a specifically tailored dataset and diversity-aware evaluation metrics.

Our findings reveal that, in contrast to the case of web search, implicit diversification methods perform better than the explicit ones for diversifying tweet search results. This contradictory result is due to the fact that most of the additional terms describing the query aspects do not appear in tweets, which are very short pieces of text. Among the implicit methods, Sy and Max-Sum are found to be the top-performers for different evaluation metrics.

The rest of the paper is organized as follows. In the next section, we present the features utilized for computing the relevance and similarity scores to be used in tweet ranking. In Section 3, we discuss how various diversification methods are adopted for ranking results in tweet search. We present our experimental setup and results in Section 4. In Section 5, we briefly review the related work in the areas of search result diversification and tweet ranking. Finally, we conclude and point to future research directions in Section 6.

2 Features for Tweet Ranking

For our purposes in this paper, we first need to specify how we compute the relevance and similarity scores for the query-tweet and tweet-tweet pairs,

respectively. In what follows, we briefly discuss the features and functions employed in our tweet ranking framework.

Similarity function. While computing the similarity of a pair of tweets, we use three types of features, namely, the pure textual content (i.e., terms), hashtags, and tweet time, as follows.

- *Content features.* While extracting the textual content, we remove the media links, urls, mention tags and hashtags in the tweets and reduce the tweet content only to a set of terms. Then, we stem the terms using JWNL (Java WordNet Library). The similarity score between two sets of terms is computed using various well-known functions from the literature. In particular, we employ the traditional Jaccard, Cosine and BM25 functions while computing the similarity of tweets. Whenever needed, IDF values are obtained over the initial retrieval results for a given query.

  Following the practice in the earlier works that employ simpler overlap-based functions in case of tweet ranking (e.g., [21]), we also define a ratio-based similarity function. This function computes the percentage of common terms in tweets t_i and t_j as shown in Eq. 1. In our preliminary analysis, we observed that Twitter users tend to use hashtags also as words to construct a full sentence as in the following example: "Ya Libnan: #Lebanon #credit #rating suffering because of #Syria #crisis"[1]. So, in the experiments, we also employ a variant of this function (denoted as Ratio-H in Section 4) that considers each hashtag as a typical term by stripping of the # sign.

$$S_W(t_i, t_j) = \frac{|Terms(t_i) \cap Terms(t_j)|}{|Terms(t_i)|} \tag{1}$$

- *Hashtag features.* For a given pair of tweets, we compute the Jaccard similarity of the set of hashtags, which is denoted as S_H.
- *Time feature.* Time similarity score between two tweets is based on the difference between their normalized timestamps (using Min-Max Normalization), and computed by Eq. 2.

$$S_T(t_i, t_j) = 1 - |t_{norm}(t_i) - t_{norm}(t_j)| \tag{2}$$

The overall tweet-tweet similarity is computed as a linear weighted function of the content similarity (S_W), hashtag similarity (S_H) and time similarity (S_T) scores, as shown in Eq. 3. In this equation, α_i represents the weight for the similarity score for each feature, where $\sum_i \alpha_i = 1$.

$$sim(t_i, t_j) = \alpha_1 S_W(t_i, t_j) + \alpha_2 S_H(t_i, t_j) + \alpha_3 S_T(t_i, t_j) \tag{3}$$

Relevance function. We compute the relevance of a query to a given tweet, i.e., $rel(q, t)$, using only term features, as hashtags and time features are not available for the queries in our dataset. In our evaluations, we consider all four functions employed for the similarity computation case, namely, Jaccard, Cosine, BM25 and Ratio functions, while computing the relevance scores.

[1] The sentence is a tweet collected for topic 7, namely "syria civil war".

3 Tweet Ranking Using Diversification Methods

Let's assume a query q that retrieves a ranked list of tweets C (where $|C| = N$) over a collection of tweets. Our goal in this work is to obtain a top-k ranking S (where $k < N$) that maximizes both the relevance and diversity among all possible size-k rankings S' of C.

To achieve our goal, we employ six different diversification methods that are, to the best of our knowledge, not applied in tweet ranking framework before. In particular, we adopt LexRank [8] and Biased LexRank [14] from text summarization domain, and MMR [3], Max-Sum [10], Max-Min [10], and xQuAD [19] methods from web search domain. All methods except xQuAD are implicit methods, as they only rely on the initial retrieval results for obtaining the top-k diversified ranking. In contrast, xQuAD is a representative of the explicit diversification paradigm, which leverages apriori information regarding the aspects (sub-topics) of the queries during the ranking. Finally, we also include the Sy method proposed in [21,22]. As far as we know, this is the only method directly applied for tweet search diversification in the literature. In what follows, we briefly summarize each of these methods.

Tweet Ranking using LexRank. LexRank [8] is a graph-based multi-document summarization approach that constructs a graph of the input sentences and ranks the sentences performing random walks. The score of the sentences, namely the score vector p, is computed by Eq. 4. In this equation, A is a square matrix with all elements being set to to $1/M$, where M is the number of sentences. As usual in a random walk, A matrix represents the probability of jumping to a random node in the graph. The pairwise similarities of the sentences are captured in the matrix B. When this algorithm converges, the sentences with the highest scores are selected to construct the summary.

$$p = [\lambda A + (1 - \lambda)B]^T p \tag{4}$$

In this work, instead of sentences, we use tweets that are in the initial retrieval set C for a given query q, and apply the same algorithm to select top-k tweets into S. Note that, some earlier works (e.g.,[20]) also employ LexRank for summarizing tweet streams; however, they do not provide an evaluation based on diversity-aware IR metrics as we do in this paper.

Tweet Ranking using Biased LexRank. Biased LexRank [14] is an extension of LexRank method, which additionally takes into account the relevance of the documents (in our case, tweets) to a given query. The computation is performed using the same formula shown in Eq. 4. However, in this case, A represents the query-tweet relevance matrix. As in the LR method, top-k tweets with the highest stationary probabilities after the convergence are selected into S.

Tweet Ranking using MMR. MMR [3] is a well-known greedy method to combine query relevance and information novelty. In MMR, the set S is initialized with the tweet that has the highest relevance to the query. In each iteration, MMR reduces the relevance score of a candidate tweet by its maximum similarity

to already selected tweets in S, and then selects the next tweet based on these discounted scores, i.e., the tweet that maximizes Eq. 5.

$$f_{MMR}(t_i) = \lambda\, rel(t_i, q) - (1 - \lambda) \max_{t_j \in S} sim(t_i, t_j) \tag{5}$$

Note that, in MMR, we involve a trade-off parameter λ to balance the relevance and diversity in the final result set.

Tweet Ranking using Max-Sum and Max-Min. We adopted two of the diversification methods proposed in [10], namely Max-Sum and Max-Min approaches that are based on the solutions for the facility dispersion problem in operations search. In the former method, the objective function aims to maximize the sum of the relevance and diversity (i.e., dissimilarity) in the final result set. This is achieved by a greedy approximation algorithm that selects a pair of tweets that maximizes Eq. 6 in each iteration.

$$f_{MaxSum}(t_i, t_j) = (1 - \lambda)(rel(t_i) + rel(t_j)) + 2\lambda(1 - sim(t_i, t_j)) \tag{6}$$

In the Max-Min method, the objective is maximizing the minimum relevance and dissimilarity of the result set. In this case, the greedy algorithm initially selects the pair of tweets that maximizes Eq. 7. Then, in each iteration, it selects the tweet that maximizes Eq. 8. As in the case of MMR, these methods employ a parameter λ for setting the trade-off between the relevance and diversity.

$$f_{MaxMin}(t_i, t_j) = 1/2(rel(t_i) + rel(t_j)) + \lambda(1 - sim(t_i, t_j)) \tag{7}$$

$$f_{MaxMin}(t_i) = \min_{t_j \in S} f_{MaxMin}(t_i, t_j) \tag{8}$$

Tweet Ranking using xQuAD. xQuAD [19] is an explicit diversification method based on the assumption that aspects of a query can be known apriori. The method aims to maximize the coverage of tweets related to the different aspects of the query and to minimize the redundancy with respect to these aspects. The greedy algorithm selects the tweet that maximizes Eq. 9 in each iteration. In this formula, $P(t_i|q)$ denotes the relevance of t_i to query q, $P(q_i|q)$ denotes the likelihood of the aspect q_i for the query q, $P(t_i|q_i)$ denotes the relevance of t_i to the query aspect q_i, and finally the product term represents the probability of q_i not being satisfied by the tweets that are already selected into S.

$$f_{xQuAD}(t_i) = (1 - \lambda)P(t_i|q) + \lambda \sum_{q_i} \left[P(q_i|q)P(t_i|q_i) \prod_{t_j \in S} (1 - P(t_j|q_i)) \right] \tag{9}$$

Tweet Ranking using Sy In a recent study [21], Tao et al. present a framework for detecting duplicate and near-duplicate tweets and define a large set of features to be employed in this context. Furthermore, they define a simple yet

effective diversification method, so-called Sy, leveraging these features. Their method scans the list of initial retrieval results, C, in a top-down fashion. For each tweet t_i, all the succeeding (lower-ranked) tweets that are near-duplicates of t_i (i.e., with a similarity score greater than a pre-defined threshold) are removed.

4 Experiments

Dataset. For our evaluations, we use the Tweets2013 corpus [22] that is specifically built for tweet search result diversification problem. The dataset includes tweets collected between February 1, 2013 and March 31, 2013. There are forty seven query topics and each topic has, on the average, 9 sub-topics.

The owners of the Tweets2013 corpus only share the tweet identification numbers, as the Twitter API licence does not allow users to share the content of the tweets. Using the provided IDs and Twitter API, we attempted to obtain top-100 and top-500 tweets for each topic. Since some of these tweets were erased or their sharing status were changed, we ended up with 81 tweets per topic for top-100 set and 403 tweets per topic for top-500 set, respectively, on the average.

In top-100 tweet collection, we observed that 80% of the tweets are not assigned to any sub-topics. In particular, there are four query topics (with ids 5, 9, 22 and 28) for which the resulting tweets are not related to any of their sub-topics. Besides, there are six more topics (with ids 7, 8, 14, 43, 46 and 47) that retrieve at most 2 relevant tweets among their top-100 results. We removed all of these topics from our query set to avoid misleading or meaningless results.

Similarly, in our top-500 tweet collection, 91% of the tweets are not assigned to any sub-topics. In this case, there are eleven topics (with ids 3, 5, 7, 9, 14, 28, 37, 43, 45, 46 and 47) for which less than 3% of the tweets in top-500 are relevant, and one additonal topic (with id 22) having no relevant tweets at all. These topics are again removed from our query set in the experiments that employ top-500 collection.

Evaluation Metrics. We evaluate diversification methods using the `ndeval` software[2] employed in TREC Diversity Tasks. We report results using three popular metrics, namely, α-nDCG [6], Precision-IA [1], and Subtopic-Recall [27] at the cut-off values of 10 and 20, as typical in the literature.

Results. We present the evaluation results for the methods adopted in this paper, namely LexRank (LR), Biased LexRank (BLR), MMR, Max-Sum, Max-Min and xQuAD, as well as the Sy method that is previously utilized for tweet diversification. We also report the performance for the initial retrieval results (i.e., without any diversification) obtained by a system employing the query-likelihood (QL) retrieval model. Note that, these initial retrieval results were provided in the Tweets2013 corpus, however we re-compute their effectiveness scores based on only those tweets that were still accessible using Twitter API, for the sake of fair comparison. Therefore, the effectiveness of the baseline QL run slightly differs from what is reported in [22].

[2] `http://trec.nist.gov/data/web10.html`

Table 1. Effectiveness of diversification methods for $N = 100$ (We denote content, hashtag and time features that are used in the similarity functions with C, H and T, respectively. Ratio-H function computes the ratio-based similarity using both terms and hashtags).

				α-nDCG		Prec-IA		ST-Recall	
Method	Rel.	Sim.	Sim. Features	@10	@20	@10	@20	@10	@20
QL	-	-	-	0.303	0.346	0.065	0.058	0.357	0.505
SY	QL	Sy	Syntactic [22]	0.339	0.378	0.080	0.068	0.401	0.529
SY	QL	Jaccard	C,H,T	**0.348**	**0.383**	**0.083**	**0.069**	**0.419**	0.542
LR	BM25	BM25	C, H, T	0.301	0.342	0.066	0.059	0.361	0.486
BLR	BM25	BM25	C, H, T	0.316	0.344	0.067	0.055	0.382	0.473
MMR	Ratio-H	Ratio-H	C	0.341	0.374	0.066	0.056	0.417	0.539
MaxSum	Ratio-H	Cosine	C, H	0.325	0.374	0.064	0.060	0.397	**0.561**
MaxMin	Ratio	Cosine	C	0.322	0.365	0.060	0.057	0.380	0.527
xQuAD	Jaccard	Jaccard	C	0.235	0.263	0.050	0.041	0.302	0.419

In our evaluations, we employed the diversification methods to compute the final top-k ranking S out of N initial retrieval results, where k is 30 and N is from {100, 500}. In Tables 1 and 2, we report the best-results (based on the α-nDCG@20 scores) for each method when N is 100 and 500, respectively. We also present the functions and features that are used for computing the relevance and similarity scores in each case[3]. For the Sy method, in addition to using the features described in Section 2, we also experimented with its best performing setup reported in a previous study, i.e., employing the syntactic feature set with the associated feature weights for computing the tweet-tweet similarity [22]. For xQuAD, following the practice in [19], we use the official query sub-topics provided in the dataset to represent an ideal scenario.

Table 1 reveals that Max-Sum and Sy are the best diversification strategies for different evaluation metrics when N is set to 100. In particular, Sy (with our features) outperforms all other methods in terms of the Prec-IA metric, whereas Max-Sum achieves the highest score for the ST-Recall@20. Note that, MMR also outperforms the Sy version that incorporates the syntactic features in [22] in terms of the ST-Recall. MMR and both versions of Sy are the best performers for α-nDCG metric and yield comparable results to each other. We also observe that BLR, the query-aware version LR, is slightly better than the original algorithm.

A surprising result that is drawn form Table 1 is that implicit diversification methods outperform xQuAD, an explicit diversification strategy, by a wide margin. This is contradictory to the findings in the case of web search result diversification, where explicit methods are usually the top-performers. For further insight on this finding, we analyzed the occurrence frequency of sub-topic terms in our tweet collection. It turns out that most of the sub-topic terms do

[3] We only report the features employed in the similarity functions, as all the relevance functions use just the content (terms) feature.

Table 2. Effectiveness of diversification methods for $N = 500$ (We denote content, hashtag and time features that are used in the similarity functions with C, H and T, respectively. Ratio-H function computes the ratio-based similarity using both terms and hashtags.)

				α-nDCG		Prec-IA		ST-Recall	
Method	Rel.	Sim.	Sim. Features	@10	@20	@10	@20	@10	@20
QL	-	-	-	0.303	0.346	0.065	0.058	0.357	0.505
SY	QL	Sy	Syntactic [22]	0.339	0.378	0.081	**0.069**	0.402	0.529
SY	QL	Jaccard	C,H,T	**0.348**	**0.382**	**0.082**	0.068	**0.419**	**0.542**
LR	BM25	BM25	C, H, T	0.302	0.341	0.066	0.059	0.361	0.480
BLR	BM25	BM25	C, H, T	0.301	0.340	0.066	0.058	0.362	0.482
MMR	Ratio-H	Ratio-H	C	0.207	0.264	0.043	0.047	0.296	0.467
MaxSum	Ratio-H	Cosine	C, H	0.223	0.287	0.049	0.053	0.311	0.483
MaxMin	Ratio	Cosine	C	0.175	0.238	0.036	0.042	0.270	0.459
xQuAD	Jaccard	Jaccard	C	0.113	0.140	0.0202	0.020	0.142	0.233

Table 3. Effectiveness of diversification methods with the syntactical features in [21,22] and for $N = 100$

	α-nDCG		Prec-IA		ST-Recall	
Method	@10	@20	@10	@20	@10	@20
LR-Syntactic	0.191	0.243	0.035	0.039	0.271	0.438
BLR-Syntactic	0.201	0.256	0.038	0.042	0.278	0.447
MMR-Syntactic	0.147	0.203	0.033	0.038	0.263	0.407
MaxSum-Syntactic	0.118	0.145	0.022	0.020	0.175	0.254
MaxMin-Syntactic	0.097	0.128	0.019	0.020	0.156	0.272
xQuAD-Syntactic	0.207	0.268	0.046	0.048	0.289	0.469

not appear in the top-100 tweets retrieved for the corresponding queries. More specifically, we find that while tweets lack only 30% of the query terms on the average, they lack 85% of the terms appearing in the sub-topics. We believe that this is due to the way sub-topics are formulated in the Tweets2013 corpus. While defining the sub-topics, human judges seem to use more general expressions that are unlikely to overlap with the terms in the actual tweets (e.g., see the example in [22] for the sub-topics of "Hillary Clinton" query). This implies that there is room for improving the performance of explicit diversification methods, by using external sources such as an ontology or query reformulations from a query log (as in [19]) for a better representation of the sub-topics.

Table 2 reveals that Sy with our similarity features is the best diversification strategy for different evaluation metrics when N is set to 500. LR and BLR scores on top-500 are similar to those using top-100 results. However, a significant decrease in the performance of the algorithms MMR, MaxSum, MaxMin and xQuAD is observed when N is increased to 500. The latter methods seem to trade-off relevance against diversity when the initial tweet set size is increased.

As a final experiment, we incorporate the syntactical features used for the Sy method into all other diversification methods while computing the similarity

scores for the tweet pairs. These syntactical features include Levenshtein Distance between tweet contents, overlap in terms, overlap in hashtags, overlap in URLs, overlap in extended URLs and length difference (please refer to [21,22] for details). In this case, query-tweet relevance scores are computed based on the content (term) overlap. In this experiment, we use the best-performing feature weights obtained via logistic regression in [21]. Our results in Table 3 reveal that these features are less useful for the diversification methods we consider in this paper; and usually degrade their performance (cf. Table 1). Note that, we only report the results for $N = 100$ for the sake of brevity.

5 Related Work

5.1 Ranking Tweets

There exists a considerable number of studies which focus on tweet ranking. Relevance to the search query is the major ranking criteria in most of the work on tweet ranking for Twitter search. Jabeur et al. [12] model the relevance of a tweet to a query by a Bayesian network that integrates a variety of features, namely micro-blogger's influence on the query topic, time and content features of tweets. In [29], the authors train machine learning models for ranking tweets against a query. Another study [13] reports that taking the hyper-links in tweet content into account improves the relevance of the retrieved results.

To the best of our knowledge, there are only two earlier studies, namely [18] and [22], that consider novelty as an additional criteria to relevance for ranking tweets. In the first study [18], an approach based on MMR [3] and clustering of tweets is proposed. However, they do not evaluate the proposed approach using diversity-aware evaluation metrics, as we do here. In the second study, Tao et al. [22] introduce Tweets2013 corpus, a data set designed for evaluating result diversification approaches for Twitter search. They also report the diversification performance of their duplicate detection framework introduced in [21] on this corpus. Note that, in this paper, we compare several other approaches to their method, Sy, in a framework that again employs Tweets2013 corpus.

Personalized tweet ranking aims to rank tweets according to the likelihood of being liked by a user. Feng et al. propose a model for personalization of Twitter stream based on the observation that a user is more interested in a tweet if she is likely to retweet it [9]. Therefore, tweets are ranked with respect to the likelihood of being re-tweeted by a user. Retweet likelihood is modeled with a graph that incorporates information from different sources such as the user's profile and interaction history.

Another recent study for personalized tweet ranking is [25]. Vosecky et al. propose a model for delivering personalized and diverse content in response to a search query. In particular, they explicitly represent both a user's and her friends' interests using topic models, and re-rank the search results based on these models. Note that, their evaluation is again based on traditional IR metrics, i.e., they do not explicitly evaluate whether the search results cover different aspects of a given query.

5.2 Diversifying Web Search Results

As web queries are inherently ambiguous and/or underspecified, diversifying the search results to cover the most probable aspects of a query among the top-ranked results (usually, top-10 or -20) arise as a popular research topic. In the *implicit diversification methods*, such query aspects are discovered from the initial retrieval results in various ways that usually involve constructing clusters [11] or topic models [5]. Since finding the optimal diversification is shown to be NP-hard [4], greedy approximation heuristics need to be employed. In this sense, a large number of implicit methods employ greedy best-first search strategy. The representative methods in this category include MMR [3], risk minimization framework proposed by Zhai et al. [28], Greedy Marginal Contribution [24] method that extends the traditional MMR, and Modern Portfolio Theory (MPT) that takes into account the variance of the relevance of the query results over different query aspects [26,17]. Gollapudi and Sharma model the result diversification problem as a bi-criteria optimization problem and then cast it to the well-known obnoxious facility dispersion problem in Operations Research [10]. In this framework, depending on the objective function, it is possible to adopt the greedy heuristics such as the Max-Sum and Max-Min approaches. In contrast, Zuccon et al. cast the diversification problem to the desirable facility dispersion problem and apply greedy local search strategy to find an approximate solution [30]. Vieira et al. also consider a semi-greedy strategy based on local search to obtain diversified query results [24].

In the *explicit diversification methods*, we assume that query aspects are known apriori, i.e., discovered from a taxonomy or query log. To this end, in one of the earliest studies, Radlinski and Dumais utilize query re-formulations [16]. In contrast, IA-Select strategy assumes the existence of a taxonomy that can be used to assign category labels to queries and retrieved results, and exploit these labels for diversification [1]. The xQuAD strategy again makes use of the query reformulations to discover the aspects and proposes a probabilistic mixture model to construct the diversified query result [19]. Dang and Croft introduce a proportionality based approach that takes into account the representation proportion of each aspect in the top-ranked query results [7]. In a recent study, Ozdemiray and Altingovde adapt score- and rank aggregation methods to the result diversification problem and show that they are both effective and efficient in comparison to the earlier methods [15]. Our work in this paper employs representative approaches from both of the implicit and explicit diversification methods to shed light their performance in the context of tweet search.

6 Conclusion

In this paper, we presented an empirical analysis of a variety of search result diversification methods adopted from the text summarization and web search domains for the task of tweet ranking. Our experiments revealed that the implicit diversification methods outperform a popular explicit method, xQuAD, due to the vocabulary gap between the official query sub-topics and tweets. Among

the implicit methods, while Sy seems to be the most promising one; there is no clear winner, and different strategies yield the best (or comparable) results for different diversity-aware evaluation metrics.

As a future work, we plan to incorporate additional features such as re-tweet counts, media links, and user popularity. We also aim to investigate the performance of explicit diversification methods with better sub-topic descriptions, and explore how such sub-topic descriptions can be automatically extracted using the clues available in a microblogging platform.

Acknowledgments. This work is partially funded by METU BAP-08-11-2013-055 project. I. S. Altingovde acknowledges the Yahoo! Faculty Research and Engagement Program.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proc. of WSDM 2009, pp. 5–14 (2009)
2. Busch, M., Gade, K., Larson, B., Lok, P., Luckenbill, S., Lin, J.: Earlybird: Real-time search at twitter. In: Proc. of ICDE 2012, pp. 1360–1369 (2012)
3. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proc.of SIGIR 1998, pp. 335–336 (1998)
4. Carterette, B.: An analysis of NP-completeness in novelty and diversity ranking. In: Azzopardi, L., Kazai, G., Robertson, S., Rüger, S., Shokouhi, M., Song, D., Yilmaz, E. (eds.) ICTIR 2009. LNCS, vol. 5766, pp. 200–211. Springer, Heidelberg (2009)
5. Carterette, B., Chandar, P.: Probabilistic models of ranking novel documents for faceted topic retrieval. In: Proc. of CIKM 2009, pp. 1287–1296 (2009)
6. Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proc. of SIGIR 2008, pp. 659–666 (2008)
7. Dang, V., Croft, W.B.: Diversity by proportionality: an election-based approach to search result diversification. In: Proc. of SIGIR 2012, pp. 65–74 (2012)
8. Erkan, G., Radev, D.R.: Lexrank: graph-based lexical centrality as salience in text summarization. J. Artif. Int. Res. 22(1), 457–479 (2004)
9. Feng, W., Wang, J.: Retweet or not?: personalized tweet re-ranking. In: Proc. of WSDM 2013, pp. 577–586 (2013)
10. Gollapudi, S., Sharma, A.: An axiomatic approach for result diversification. In: Proc. of WWW 2009, pp. 381–390 (2009)
11. He, J., Meij, E., de Rijke, M.: Result diversification based on query-specific cluster ranking. JASIST 62(3), 550–571 (2011)
12. Jabeur, L.B., Tamine, L., Boughanem, M.: Uprising microblogs: a bayesian network retrieval model for tweet search. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing, SAC 2012, pp. 943–948. ACM (2012)
13. McCreadie, R., Macdonald, C.: Relevance in microblogs: Enhancing tweet retrieval using hyperlinked documents. In: Proc. of the 10th Conference on Open Research Areas in Information Retrieval, OAIR 2013, pp. 189–196 (2013)

14. Otterbacher, J., Erkan, G., Radev, D.R.: Biased lexrank: Passage retrieval using random walks with question-based priors. Inf. Process. Manage. 45(1), 42–54 (2009)
15. Ozdemiray, A.M., Altingovde, I.S.: Explicit search result diversification using score and rank aggregation methods. In: JASIST (in press)
16. Radlinski, F., Dumais, S.T.: Improving personalized web search using result diversification. In: Proc. of SIGIR 2006, pp. 691–692 (2006)
17. Rafiei, D., Bharat, K., Shukla, A.: Diversifying web search results. In: Proc. of WWW 2010, pp. 781–790 (2010)
18. Rodriguez Perez, J.A., Moshfeghi, Y., Jose, J.M.: On using inter-document relations in microblog retrieval. In: Proc. of WWW 2013, pp. 75–76 (2013)
19. Santos, R.L., Macdonald, C., Ounis, I.: Exploiting query reformulations for web search result diversification. In: Proc. of WWW 2010, pp. 881–890 (2010)
20. Sharifi, B., Inouye, D., Kalita, J.K.: Summarization of twitter microblogs. Comput. J. 57(3), 378–402 (2014)
21. Tao, K., Abel, F., Hauff, C., Houben, G.-J., Gadiraju, U.: Groundhog day: Near-duplicate detection on twitter. In: Proc. of WWW 2013, pp. 1273–1284 (2013)
22. Tao, K., Hauff, C., Houben, G.-J.: Building a microblog corpus for search result diversification. In: Banchs, R.E., Silvestri, F., Liu, T.-Y., Zhang, M., Gao, S., Lang, J. (eds.) AIRS 2013. LNCS, vol. 8281, pp. 251–262. Springer, Heidelberg (2013)
23. Teevan, J., Ramage, D., Morris, M.R.: #twittersearch: a comparison of microblog search and web search. In: Proc. of WSDM 2011, pp. 35–44 (2011)
24. Vieira, M.R., Razente, H.L., Barioni, M.C.N., Hadjieleftheriou, M., Srivastava, D., C. T. Jr., Tsotras, V.J.: On query result diversification. In: Proc. of ICDE 2011, pp. 1163–1174 (2011)
25. Vosecky, J., Leung, K.W.-T., Ng, W.: Collaborative personalized twitter search with topic-language models. In: Proc. of SIGIR 2014, pp. 53–62 (2014)
26. Wang, J., Zhu, J.: Portfolio theory of information retrieval. In: Proc. of SIGIR 2009, pp. 115–122 (2009)
27. Zhai, C., Cohen, W.W., Lafferty, J.D.: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: Proc. of SIGIR 2003, pp. 10–17 (2003)
28. Zhai, C., Lafferty, J.D.: A risk minimization framework for information retrieval. Inf. Process. Manage. 42(1), 31–55 (2006)
29. Zhang, X., He, B., Luo, T., Li, B.: Query-biased learning to rank for real-time twitter search. In: Proc. of CIKM 2012, pp. 1915–1919 (2012)
30. Zuccon, G., Azzopardi, L., Zhang, D., Wang, J.: Top-k retrieval using facility location analysis. In: Baeza-Yates, R., de Vries, A.P., Zaragoza, H., Cambazoglu, B.B., Murdock, V., Lempel, R., Silvestri, F. (eds.) ECIR 2012. LNCS, vol. 7224, pp. 305–316. Springer, Heidelberg (2012)