Evaluating Ranking Diversity and Summarization in Microblogs using Hashtags

David Fisher[1], Ashish Jain[1], Mostafa Keikha[1], W. B. Croft[1], and Nedim Lipka[2]

[1] University of Massachusetts, Amherst MA 01003 USA
{dfisher,ashishjain,keikham,croft}@cs.umass.edu
[2] Systems Technology Lab Adobe Systems, Inc. San Jose, CA 95110 USA
lipka@adobe.com

Abstract. Diversification techniques for web search have recently been developed that assume that, for each query, there is a set of underlying aspects or subtopics that address specific user intents. These techniques attempt to balance the relevance of the retrieved documents with the coverage of the aspects. Evaluation of diversification techniques requires some way of defining a set of aspects for each test query and a "gold standard" assignment of documents to aspects. This has made the study of diversification difficult for new data such as microblogs. A related task, keyword-based summarization, is important for microblogs but also has problems in evaluation. In this paper, we describe an approach to evaluating ranking diversity and summarization in microblogs by assuming hashtags correspond to subtopics. We show the viability of this approach to evaluation, and validate the assumption that hashtags are subtopics. The results show that, despite the differences in content, the best techniques for search diversification with microblogs are the same as with web pages. The summarization results confirm that the DSPapprox technique is effective and that phrase-based summarization techniques perform somewhat worse than single words in terms of covering the underlying aspects.

Keywords: microblog diversity, summarization, evaluation

1 Introduction

The motivation for search diversification is that, given a query, there may be multiple user intents related to that query and an effective ranking should try to cover each of those intents by including a diverse range of documents. To do this, search diversification techniques such as xQuAD [13] and the proportionality model [6] explicitly model the underlying aspects or subtopics for a query and select documents based on a combination of their relevance to the original query and relevance to the aspects. The aspect descriptions have been generated using a variety of methods in previous research. Most approaches have relied on manual creation, either as a list of topics [13], [6], a larger taxonomy from which the query aspects can be inferred [1], or a list of topics obtained directly from

commercial search engines [13], [6]. More recently, an approach to automatic aspect or topic generation based on the vocabulary of the retrieved documents has been described [7]. This approach, called term-level diversification, has significantly outperformed other methods for automatic generation and has comparable results to the best methods using manually created aspects. Evaluating a diversification method, however, still requires a "gold standard" categorization of documents and aspects for each query, regardless of whether automatic or manually generated aspects are used to create the diversified ranked list. This creates a significant barrier for diversification experiments with new collections such as microblogs or other social media.

In this paper, we propose an approach to evaluating ranking diversity specifically for microblogs or any social media that uses hashtags. Our hypothesis is that the hashtags can be treated as manually assigned subtopics or aspects associated with the microblog post (i.e., the gold standard). A simplified description of our process is that we use the hashtags that are highly associated with the posts retrieved by a query as aspects for that query. A list of "stoptags" is used to filter the common hashtags that are used frequently across all topics. Search diversification is then performed using only the text part of the posts and the resulting ranking is evaluated using the assignment of hashtags to posts. We describe experiments and examples that show that this approach produces reasonable results, and compare state-of-the-art diversification techniques for microblog data.

We evaluate the assumption that hashtags provide subtopics with a manual evaluation for our experimental collections. This evaluation shows that there is reasonable agreement concerning topicality across multiple judges. It also shows that between one half and three quarters of the hashtags are considered subtopical by at least one judge.

One task where diversity is particularly important is microblog summarization. In this task, the goal is to summarize the major discussion themes or subtopics in a microblog stream for a given query topic. Figure 1 shows a screenshot of a demonstration system developed at Adobe that shows the most important phrases and hashtags from a Twitter stream for the query "adobe". Users can browse the list of topics and select them to look at specific groups of tweets.

There has been previous research on techniques to generate sentence or keyword-based microblog summaries [11], [14, 15], but evaluating these techniques has been a major problem. The evaluation of summarization techniques has often relied on comparison to the gold standard manual summary and, based on this, we suggest an evaluation technique for microblog summarization that relies on the hashtags as the manual summary. Since hashtags cannot be compared directly to keywords or phrases, we instead propose to evaluate summarization techniques by the diversity or coverage of the ranking they generate in conjunction with a diversification technique. Although this is an extrinsic evaluation method, it does make direct use of the manually generated hashtag summary. We describe experiments comparing various summarization techniques using this evaluation approach.

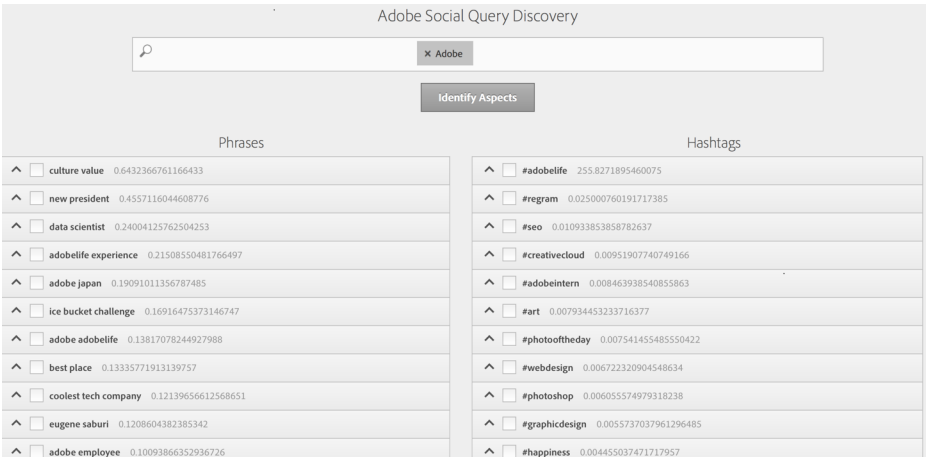

Fig. 1. Popular phrases and hashtags in Twitter stream for query "adobe".

In the next section, we describe the diversification and summarization techniques that we are studying. The term-based diversity approach [7] is emphasized since it is the basis of our approach. In addition, we describe how the diversification and summarization tasks have been evaluated and the metrics that have been used. In section 3, we detail the process we propose for evaluating ranking diversity and evaluating keyword-based summaries using hashtags. We detail the process for manual evaluation of the hashtags as subtopics in section 3.1. The test collections we use are covered in section 4. Section 5 presents and discusses the results. We conclude the paper in section 6 with a summary of the results and a comparison to recent related work.

2 Related Work

2.1 Diversification techniques

There are two notions of diversity in the current literature: diversity by *redundancy* and by *proportionality*. Documents at any position in the result list that provide the same information as those at earlier ranks are considered redundant. A ranking is more diverse if it contains less redundancy, or equivalently, more novelty. Probably the best-known of the techniques that are based on redundancy is xQuAD [13], which specifies a greedy algorithm to sequentially select documents with minimal redundancy, measured by how much the new document covers the query subtopics that have not been well covered by those chosen earlier. On the other hand, a proportional ranking of documents with respect to a subtopic popularity distribution is a ranking in which the number of documents on each subtopic is proportional to its popularity [6]. By this definition, perfectly proportional search results would naturally be diverse. The main algorithm in this class is PM-2 [6], which selects documents in a similar greedy fashion, except

that it maximizes proportionality using the Sainte-Laguë formula. In this paper, we compare xQuAD and PM-2 used for term-level diversification in microblogs. We now describe the difference between topic-level diversification and term-level diversification.

Topic Level Diversification Let q indicate a user query and $T = \{t_1, t_2, ..., t_n\}$ indicate the set of topics for q. Let $W = \{w_1, w_2, ..., w_n\}$ denote the weights for each of the topics $t_i \in T$ (we refer to the subtopics for a query as topics for brevity). These weights can be interpreted as the importance [13] or popularity [6] depending on the diversification technique. In addition, let $R = \{d_1, d_2, ..., d_m\}$ indicate a ranked list of documents initially retrieved for q and $P(d|t)$ denote some probabilistic estimate of d's relevance to a topic t. The task of topic level diversification is to select a subset of R using $\{T, W, P(d|t)\}$ to form a diverse ranked list S of size k.

The *xQuAD* framework promotes diverse rankings of documents by penalizing redundancy at every rank. It does so by greedily selecting documents in R to put into S. At each step, it selects a highly-ranked document that is most different to those previously selected (thus minimizing redundancy):

$$d^* = \arg\max_{d_j \in R} \ (1 - \lambda) \times P(d_j|q) + \lambda \times D(d_j, S) \tag{1}$$

xQuAD measures the difference between documents by the topics they cover. It defines p_i to be the "portion" of the topic t_i that has not been covered by documents in S:

$$p_i = \prod_{d_j \in S} (1 - P(d_j|t_i)) \tag{2}$$

Higher p_i indicates that most of the documents in S are not relevant to t_i. As such, t_i is less substantially covered and it should have higher "priority" in getting more documents. With this, $D(d_j, S)$ is calculated as follows:

$$D(d_j, S) = \sum_{t_i \in T} w_i \times P(d_j|t_i) \times p_i \tag{3}$$

which means the novelty of a document is its ability to cover the topics that need covering (i.e. higher p_i) weighted by the importance of the topics w_i.

The *proportionality model PM-2* [6] is a probabilistic adaptation of the Sainte-Laguë method for assigning seats to members of competing political parties such that the number of seats for each party is proportional to the votes they receive. PM-2 starts with a ranked list S with k empty seats. For each of these seats, it computes the quotient qt_i for each topic t_i following the Sainte-Laguë formula:

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

According the the Sainte-Laguë method, this seat should be awarded to the topic with the largest quotient in order to best maintain the proportionality of

the list. Therefore, PM-2 assigns the current seat to the topic t_{i^*} with the largest quotient. The document to fill this seat is the one that is not only relevant to t_{i^*} but to other topics as well:

$$d^* = \arg\max_{d_j \in R} \ \lambda \times qt_{i^*} \times P(d_j|t_{i^*}) + (1-\lambda) \sum_{i \neq i^*} qt_i \times P(d_j|t_i) \tag{5}$$

After the document d^* is selected, PM-2 increases the "portion" of seats occupied by each of the topics t_i by its normalized relevance to d^*:

$$s_i = s_i + \frac{P(d^*|t_i)}{\sum_{t_j \in T} P(d^*|t_j)} \tag{6}$$

This process repeats until we get k documents for S or we are out of candidate documents. The order in which each document is put into S determines its ranking.

Term Level Diversification Diversification at the term level is very similar to the topic level. Let $t_i = \{t_i^1, t_i^2, ..., t_i^{|t_i|}\}$ be the set of vocabulary terms (words, phrases) for topic t_i. Instead of diversifying R using the set of topics $T = \{t_1, t_2, ..., t_n\}$, we perform diversification using

$$T' = \{t_i^1, t_i^2, ..., t_i^{|t_i|}, ..., t_n^1, t_n^2, ..., t_n^{|t_n|}\}$$

in effect treating each t_i^j as a topic. Previous research [7] using TREC web collections has shown that term-level diversification produces superior results compared to topic-based diversification. The main issue in this approach is generating the terms associated with a query, which has been done using the summarization techniques discussed in section 2.3.

2.2 Evaluating Diversification

Diversification techniques are compared using several standard metrics that have been used in the official evaluation of the diversity tasks at TREC. In this paper, we use α-NDCG [4]. We also calculated other measures, such as ERR-IA (a variant of ERR [3]), but found that these additional measures did not reveal any different trends. These metrics penalize redundancy at each position in the ranked list based on how much of that information the user has already seen from documents at earlier ranks. As we mentioned previously, in TREC evaluations, the metrics depend on a gold-standard assignment of topics to documents that is provided. In the experiments with microblog data reported in this paper, we use hashtags as an alternative. We compute all of these measures using the top 20 documents retrieved by each model to be consistent with official TREC evaluation.

Diversification can potentially have an impact on the overall effectiveness of the ranking. Therefore, in addition to all diversity measures above, we report results using two standard relevance-based metrics for retrieval: NDCG and Precision at rank 10. For the microblog data, we again use hashtags in place of relevance judgments, as described in section 3.

2.3 Summarization Techniques

For the microblog summarization task, we will be using a multi-document keyword-based summarization technique. That is, we will be summarizing a group of documents (posts retrieved by the query) using a list of keywords or phrases. A typical simple algorithm would be to list the words or phrases with the highest tf.idf weights. More complex algorithms select summary terms that maximize some combination of tf.idf with other features.

In our experiments, we use the DSPapprox algorithm [9], which maximizes a combination of two features called *predictiveness* and *topicality*. In this algorithm, terms are considered *vocabulary terms* if they (1) appear in at least two documents, (2) have at least two characters and (3) are not numbers. In our experiments, we use various types of terms: words, capitalized nouns, noun phrases, and combinations of these categories. All vocabulary terms that co-occur with any of the query terms within a proximity window of size w are selected as *topic terms*. We then compute topicality and predictiveness. The topicality of a term measures how informative it is at describing the set of documents being summarized (in this case, the top ranked posts). To compute topicality, a relevance model $P_R(t|q)$ [8] is first estimated from the initial set of documents R:

$$P_R(t|q) = \sum_{d_i \in R} P(t|d_i)P(d_i|q) \tag{7}$$

where $P(t|d)$ is the probability that d_i generates the term t and $P(d_i|q)$ is relevance of d_i to the query. The topicality $TP(t)$ of a term t is estimated as its contribution to the KL divergence between this relevance model and the language model for the entire collection:

$$TP(t) = P_R(t|q) log_2 \frac{P_R(t|q)}{P_c(t)} \tag{8}$$

It is equivalently t's contribution to the clarity score of the query q [17].

Predictiveness, on the other hand, measures how much the occurrence of a term predicts the occurrences of others. Let $P_w(t|v)$ indicate the probability that a term t occurs within a window of size w of another term v and C_t indicate the set all such v. The predictiveness of t is estimated as follows:

$$PR(t) = \frac{1}{Z} \sum_{v \in C_t} P_w(t|v) \tag{9}$$

where Z is the hierarchy level specific normalization factor. In our case, we set it to the size of the vocabulary.

The DSPapprox algorithm iteratively selects terms from the candidate topic term set T. The utility of each term is the product of its topicality and predictiveness. At each step, the algorithm selects the topic term $t^* \in T$ with maximum utility. Then, it decreases the predictiveness of other topic terms that predict the same vocabulary. This ensures that topic terms that cover the uncovered part of the vocabulary will emerge for selection in the next iteration. The algorithm stops once the utility of all candidate topic terms reaches 0, indicating that all vocabulary has been covered.

2.4 Evaluating Summarization

Sentence-based summaries are often evaluated by comparison to a gold-standard manual summary using measures such as ROUGE. A recent paper by Mackie et al [10] compares this and other summarization metrics for microblogs. The experiments they describe were limited by the absence of gold-standard summaries in their test collections. As an alternative, we are proposing to use hashtags as the gold-standard manual annotation of topics for a set of microblog posts. We cannot, however, simply use overlap measures for the list of keywords and the list of hashtags, since many of them will not match even if they are strongly related. Instead, we propose to compare the list of keywords generated by the summarization technique to the list of hashtags associated with the posts by measuring the diversity of a ranked list generated using the keywords and evaluated using hashtags. We explain this process in more detail in the next section.

3 Evaluating Diversity and Summarization using Hashtags

Hashtags can be viewed as a means for users to categorize their posts, independent of the rest of the post's content. In a microblog, there are a large number of posts that are personal, trivial, crude, or generally lacking content that is meaningful for other people. Similarly, there are many personal, crude, and content-free hashtags used in those posts. For posts with more content, however, hashtags are often used to summarize the main topic and subtopics related to the post. When tags become known to other people, they are used to categorize new posts. For example, in the example in Figure 1, some of the frequent hashtags associated with the query "adobe" are "#adobelife","#creativecloud", "#photooftheday", and "#webdesign". In our research, we assume that the hashtags, after removing very frequent, content-free tags from a stoplist, are equivalent to the manually assigned subtopics or aspects used in diversification. We then use those tags to evaluate diversification algorithms that are based on the text content of the posts, ignoring the tags. In more detail, the process we used is as follows.

1. Create a stoplist of hashtags (the "stoptags") by manual selection from the top 500 most frequent tags in the microblog collection
2. Generate a set of queries by manual selection from a list of frequent hashtags not on the stoplist. A limitation of our approach is that it requires hashtags, which many posts do not use. For this reason, we chose popular tags of general interest such as "#obamacare" and "#iran". Text queries were created from these tags, resulting in queries such as "obamacare" and "iran".
3. For each query, find the top 20 associated hashtags. This was done by simply by identifying the tags with the highest co-occurrence frequency with the hashtag that was the basis of the query. These are the hashtags used as the list of aspects for the query.

4. Run various types of ranking and diversity ranking experiments using the text corresponding to the query tag as the query and ignoring the hashtags identified in the previous step.
5. Evaluate the results using diversity evaluation methods, where a post is considered relevant to a query if it contains the main query tag, and relevant to an aspect if it contains *both* the query tag and the identified aspect tag. We examined an alternative definition where a post is relevant to an aspect if it contains *only* the aspect tag and found that this made no difference to the results.

In step 3, the diversification techniques used in our experiments are xQuAD and PM-2 using term-based diversity. Our baselines for comparison are the query-likelihood ranking [5] and the pseudo-relevance feedback technique RM3 [5]. Evaluating the summarization techniques involves comparing different methods for generating the vocabulary for the term-based diversity algorithm. We used the top weighted terms from pseudo-relevance feedback as a baseline approach to generate the aspect vocabulary. Then we compared different types of vocabulary terms in the DSPapprox algorithm.

3.1 Evaluating Hashtags as Subtopics

Recent research [16] endeavored to build a microblog diversity corpus. Queries were derived from temporally topical news events. Subtopics for each query were manually specified by individual human annotators. We have performed an analogous task, manually selecting queries from the temporally relevant hashtags then automatically generating the subtopics from associated hashtags.

We began with the assumption that these hashtags could be used as a surrogate for manually assigned subtopics. In order to test that assumption, and that the subtopics are reasonable, it was necessary to perform a manual evaluation.

We evaluated these subtopics using multiple human judges. The judges were three software engineers, all with extensive experience in information retrieval. Judges were asked to judge only whether a hashtag was a subtopic of the query, rather than the broader classification of aspect. The judging process is as follows.

Two of the judges were provided with the original query paired with each of the top 20 hashtags that constitute the subtopic set. Each topic subtopic pair was judged independently of the other pairs. The hashtags were classified as either being a subtopic (1) or not (0). The third judge was provided with the same pairs as the first two judges, and additionally provided the full text of the top 20 tweets for each pair. Judging again was binary.

The 3 sets of judgments were combined by adding the score assigned by each judge. A score of 0 indicates none of the judges thought the pair contained a subtopic. A score of 3 indicates that all of the judges thought the pair contained a subtopic. This metric allows for both liberal and conservative evaluation.

4 Test Collections

We used two collections of Twitter posts for this research. The first is a collection of 3.4 million tweets from November 2011 that we have used in other research.

Table 1. Example queries

Collection	Queries				
2011	wikileaks	terrorism	obamacare	illegals	iranelection
2013	black friday	ipad	ashes	iraq	thanksgiving

We will refer to this as the 2011 collection. The second is a collection of approximately 50 million English-language tweets from November 2013 that is available at the Internet Archive[3]. We refer to this as the 2013 collection. For the 2011 collection, we created 67 queries based on hashtags, as described in section 3. For the 2013 collection, we created 45 queries, with some examples in Table 1.

Table 2. Diversity evaluation using α-NDCG@20. Bolded indicates a significant difference to the baselines (two-tailed t-test). † indicates significantly better than xQuAD.

Collection	QL	RM-3	xQuAD	PM-2
2011	0.211	0.236	**0.321**	**0.331**†
2013	0.189	0.203	0.220	**0.254**†

5 Results

The first set of results compares microblog ranking based on diversity techniques to baselines, evaluated using diversity metrics. The two non-diversified baselines we used were query likelihood ranking (QL) and ranking using pseudo-relevance feedback (RM3). We also considered using maximal marginal relevance [2] as a baseline but since it performed substantially worse than query likelihood (e.g., 0.19 vs. 0.21 α-NDCG on the 2011 collection), we removed it from the comparison. Table 2 shows the comparison of QL, RM3, xQuAD, and PM-2 on both collections using α-NDCG at rank 20 (other measures, as we mentioned, did not change the comparison). For both xQuAD and PM-2, the vocabulary used for the term-based diversity was the top 20 words ranked by DSPapprox. The input to DSPapprox was the top 50 posts in the QL ranking.

Our results in general show that hashtags are a viable method of evaluating diverse retrieval and summarization for social media. Both the numbers obtained for the evaluation metrics and the relative rankings of methods are similar to those obtained from manual aspect judging. This was the main point of this paper, but the specific results are also interesting and worth discussing in detail. These results show that the diversity-based approaches can definitely improve effectiveness based on a diversity metric. The PM-2 method was the most effective. We also compared the performance of these techniques on the large 2013 collection using the relevance-based measures NDCG and Precision at rank 10. In this experiment, the query hashtag is used as a relevance annotation. Table 3 shows that the PM-2 diversity technique improves performance based on the relevance metrics as well as the diversity metric. These are very similar results to those obtained with TREC web track diversity task data [7].

An obvious activity would be to compare these results to those from manual annotations. However, an aspect-annotated microblog collection has not been

[3] https://archive.org/details/twitterstream

Table 3. Relevance evaluation of ranking methods on the 2013 collection.

Method	NDCG@10	P@10
QL	0.329	0.324
RM3	0.334	0.324
xQuAD	0.374	0.364
PM-2	**0.433**†	**0.411**†

Table 4. Comparison of summarization methods based on RM3 and DSP for 2011.

Vocabulary	RM3 summary	DSP summary
Word	0.318	**0.331**
Phrase	0.286	0.297
Phrase + Word	0.284	**0.309**
Phrase + Capital Noun	0.295	**0.316**

Table 5. Comparison of summarization methods based on RM3 and DSP for 2013.

Vocabulary	RM3 summary	DSP summary
Word	0.236	**0.254**
Phrase	0.226	0.216
Phrase + Word	0.226	0.226
Phrase + Capital Noun	0.220	0.231

available until very recently. Ozsoy et al [12] presented work on diversifying microblog ranking based on a manually annotated corpus. The interesting thing about their results is that they found that xQuAD was much less effective than the QL baseline and that MMR was more effective than the baseline. The most effective method relied on duplicate tweet detection. These results are quite different to ours, and to previous web-based results. Finding the reasons for the differences between these two approaches to evaluation, and which is more accurate, is important and we plan to study this in future work. As an alternative, in this paper, we evaluated the validity of the hypothesis that hashtags are subtopics in section 5.1.

Tables 4 and 5 show the α-NDCG@20 results for different methods of summarizing the topical vocabulary of the top ranked posts, for both collections. Phrases were identified using POS tagging[4]. These results show that the DSPapprox method can be more effective than a method that summarizes using the top weighted terms, especially on the 2011 corpus. They also show that the word-based summaries give better coverage of the related "subtopics" than the phrase-based summaries. This was also the result with web data [7]. Our demonstration system, however, uses phrases for summaries since they are easier for people to understand than single words for many queries. The diversity measure, while an important part of summarization effectiveness, does not represent all views of effectiveness. This diversity measure can provide an alternative perspective to complement the overlap measures described in [10].

[4] http://www.ark.cs.cmu.edu/TweetNLP/

Table 6. Subtopic judgment scores

Judgment	f 2011	f 2013	Total
0	719	224	943
1	288	236	524
2	273	125	378
3	395	207	602
Total	1675	792	2467

Table 7. Subtopic judgment scores, f<100

Judgment	2011	2013	Total
0	14	38	52
1	9	57	66
2	0	19	19
3	1	28	29
Total	25	132	157

Table 8. Not subtopics, but related

Collection	Topic	Proposed Subtopic	Actual Relation
2011	BBC	CNN	Same class (Company)
2011	Toyota	Honda	Same class (Company)
2011	Qatar	Brazil	Same class (Country)
2013	iPhone	iPad	Competing product
2013	iPadgame	iphone5	Competing platform
2013	China	Obama	Politically related

5.1 Evaluation of Subtopics

Using the process described in Section 3.1, we evaluated the generated hashtags with respect to whether or not they constitute reasonable subtopics for our queries. We also consider, anecdotally, those cases where a hashtag could be considered an aspect of the query that is not a subtopic.

Table 6 shows that more than half of the hashtags for the 2011 collection, and roughly three quarters for the 2013 collection, were considered to be subtopics using the most liberal interpretation, at least a single judge said yes. Looking at the most conservative evaluation, where all three judges agreed, approximately one quarter was judged subtopical for both the 2011 and 2013 collections.

Table 7 shows that hashtags that occur less than 100 times tend to be less likely to be considered subtopics. This is especially true for the consensus judgment case. The effect is more pronounced in the 2011 collection.

Table 8 shows some examples of where a relationship exists between the topic and subtopic hashtags, but that relationship is not within the topic hierarchy. The majority are tags for similar entities of the same class, as seen in the first four entries. While not subtopics per se, the relations shown by many of the pairings could plausibly be used to define aspects with respect to diversifying search results.

6 Conclusions

In this paper, we have shown that hashtags are a viable means of evaluating diversification and summarization in microblogs. Using hashtags as a substitute

for manually identified aspects enables a number of large-scale evaluations to be done. Manual evaluation of the hashtags as subtopics shows the hashtags produce a reasonable set of aspects for diversification, with agreement across multiple judges for almost half of the hashtags. Our experimental results show that the PM-2 term-based diversification method is the most effective, which is a similar result to that obtained with web data, and the DSPapprox summarization technique can be more effective than using the highest weighted terms.

7 Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by an award from Adobe Systems, Inc., and in part by NSF grant #CNS-1405829. 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.

References

1. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of WSDM*, pages 5-14, 2009.
2. J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings SIGIR*, pages 335-336, 1998.
3. O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of CIKM*, pages 621-630, 2009.
4. C.L.A. Clarke, M. Kolla, G.V. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR*, pages 659-666, 2008.
5. W.B. Croft, D. Metzler, and T. Strohman. Search Engines: Information Retrieval in Practice. *Addison-Wesley*, 2009.
6. V. Dang and W.B. Croft. Diversity by proportionality: An election-based approach to search result diversification. In *Proceedings of SIGIR*, pages 65-74, 2012.
7. V. Dang and W.B. Croft. Term level search result diversification. In *Proceedings of SIGIR*, pages 603-612, 2013.
8. V. Lavrenko and W.B. Croft. Relevance-Based Language Models. In *Proceedings of SIGIR*, pages 120–127, 2001.
9. D. Lawrie, W.B. Croft, and A. Rosenberg. Finding topic words for hierarchical summarization. In *Proceedings of SIGIR*, pages 349–357, 2001.
10. S. Mackie, R. McCreadie, C. Macdonald, and I. Ounis. On choosing an effective automatic evaluation metric for microblog summarisation. In *Proceedings of the 5th Information Interaction in Context Symposium*, pages 115-124, 2014.
11. B. O'Connor, M. Krieger, and D.Ahn. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In *Proceedings of ICWSM*, pages 384-385, 2010.
12. M. Ozsoy, K. Onal, and I. Altingovde. Result diversification for tweet search. In *Proceedings of Web Information Systems Engineering (WISE)*, pages 78-89, 2014.
13. R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of WWW*, pages 881-890, 2010.

14. B. Sharifi, M. Hutton, and J. Kalita. Summarizing microblogs automatically. In *Proceedings of Human Language Technologies*, pages 685-688, 2010.
15. D. Spina, E. Meij, M. de Rijke, A. Oghina, M. T. Bui, and M. Breuss. Identifying entity aspects in microblog posts. In *Proceedings of SIGIR*, pages 1089-1090. 2012.
16. K. Tao, C. Hauff, and G. Houben Building a microblog corpus for search result diversification. In *Proceedings of AIRS*, pages 251-262, 2013.
17. S Cronen-Townsend, Y. Zhou and W.B. Croft. Predicting Query Quality. In *Proceedings of SIGIR*, pages 299–306, 2002.