Building a Microblog Corpus for Search Result Diversification

Ke Tao, Claudia Hauff, Geert-Jan Houben

TU Delft, Web Information Systems, Delft, the Netherlands
{k.tao,c.hauff,g.j.p.m.houben}@tudelft.nl

Abstract. Queries that users pose to search engines are often ambiguous - either because different users express different query intents with the same query terms or because the query is underspecified and it is unclear which aspect of a particular query the user is interested in. In the Web search setting, search result diversification, whose goal is the creation of a search result ranking covering a range of query intents or aspects of a single topic respectively, has been shown in recent years to be an effective strategy to satisfy search engine users. We hypothesize that such a strategy will also be beneficial for search on microblogging platforms. Currently, progress in this direction is limited due to the lack of a microblog-based diversification corpus. In this paper we address this shortcoming and present our work on creating such a corpus. We are able to show that this corpus fulfils a number of diversification criteria as described in the literature. Initial search and retrieval experiments evaluating the benefits of de-duplication in the diversification setting are also reported.

1 Introduction

Queries that users pose to search engines are often ambiguous - either because different users express different query intents with the same query terms or because the query is underspecified and it is unclear which aspect of a particular query the user is interested in. Search result diversification, whose goal is the creation of a search result ranking covering a range of query intents or aspects of a single topic respectively, has been shown in recent years to be an effective strategy to satisfy search engine users in those circumstances. Instead of a single query intent or a limited number of aspects, search result rankings now cover a set of intents and wide variety of aspects. In 2009, with the introduction of the diversity search task at TREC [1], a large increase in research efforts could be observed, e.g. [2–5].

Recent research [6], comparing users' query behaviour on microblogging platforms such as Twitter and the Web has shown that Web search queries are on average longer than Twitter queries. This is not surprising, as each Twitter message (tweet) is limited to 140 characters and a longer query might remove too many potentially relevant tweets from the result set. Considering the success of

diversity in Web search, we believe that it is an even more important technology on microblogging platforms due to the shortness of the queries.

However, to our knowledge, no publicly available microblogging data set (i.e. a corpus and a set of topics with subtopic-based relevance judgments) exists as of yet. In order to further the work on diversity in the microblog setting, we created such a corpus[1] and describe it here.

Specifically, in this paper we make the following contributions: (i) we present a methodology for microblog-based corpus creation, (ii) we create such a corpus, and, (iii) conduct an analysis on its validity for diversity experiments. In the second part of the paper we turn to the question of (iv) how to improve search and retrieval in the diversity setting by evaluating the recently introduced de-duplication approach to microblogging streams [7].

2 Related Work

Users of (Web) serach engines typically employ short keyword-based queries to expresss their information needs. These queries are often underspecified or ambiguous to some extent [8]. Different users who pose exactly the same query may have very different query intents. In order to satisfy a wide range of users, search results diversification was proposed [9].

On the Web, researchers have been studying the diversification problem mostly based on two considerations: novelty and facet coverage. To increase novelty, mazimizing the marginal relevance while adding documents to the search results [11, 12] has been proposed. Later studies have focused on how to maximize the coverage of different facets [2] of a given query. Furthermore, there are works that consider a hybrid solution to combine benefits from both novelty-based and coverage-based approaches [13, 3].

In order to evaluate the effectiveness of search result diversification, different evaluation measures have been proposed. A number of them [10, 14–16] have been employed at the Diversity Task [1] of the Text REtrieval Conference (TREC), which ran between 2009 and 2012.

Given the difference [6] in querying behavior on the Web and microblogging sites, we hypothesize that the diversification problem is more challenging in the latter case due to the reduced length of the queries. Tao et al. [7] recently proposed a framework to detect (near-)duplicate messages on Twitter and explored its performance as a search result diversification tool on microblogging sites [7]. The approach can be categorized as novelty-based since it exploits the dependency between documents in the initial result ranking. The evaluation though was limited due to the lack of an explicit diversity microblogging corpus (i.e. a corpus with topics and subtopics as well as relevance judgments on the subtopic level). In this paper, we now tackle this very issue. We describe our methodology for the creation of a Twitter-based diversity corpus and investigate its properties. Finally, we also employ Tao et al.'s framework [7] and explore its effectiveness on this newly developed data set.

[1] The corpus is publicly available at `http://wis.ewi.tudelft.nl/airs2013/`.

3 Methodology: Creating a Diversity Corpus

We collected tweets from the public Twitter stream between February 1, 2013 and March 31, 2013 - the dates were chosen to coincide with the time interval of the TREC Microblog 2013 track[2].

After the crawl, in order to create topics, one of this paper's authors (*Annotator 2*) consulted Wikipedia's *Current Events Portal*[3] for the months February and March 2013 and selected fifty news events. We hypothesized that only topics with enough importance and more than local interests are mentioned here and thus, it is likely that our Twitter stream does contain some tweets which are pertinent to these topics. Another advantage of this approach is that we were able to also investigate the importance of time - we picked topics which are evenly distributed across the two-month time span.

Having defined the documents and topics, two decisions need to be made: (i) how to derive the subtopics for each topic, and, (ii) how to create a pool of documents to judge for each topic (and corresponding set of subtopics). Previous benchmarks have developed different approaches for (i): e.g. to derive subtopics post-hoc, i.e. after the pool of documents to judge has been created or to rely on external sources such as query logs to determine the different interpretations and/or aspects of a topic. With respect to (ii), the setup followed by virtually all benchmarks is to create a pool of documents to judge based on the top retrieved documents by the benchmark participants, the idea being that a large set of diverse retrieval systems will retrieve a diverse set of documents for judging.

Since in our work we do not have access to a wide variety of retrieval systems to create the pool, we had to opt for a different approach: one of this paper's authors (*Annotator 1*) *manually* created complex Indri[4] queries for each topic topics. We consider this approach a valid alternative to the pool-based approach, as in this way we still retrieve a set of diverse documents. A number of examples are shown in Table 1. The Indri query language allows us to define, among others, synonymous terms within $< .. >$ as well as exact phrase matches with #1(...). The *#combine* operator joins the different concepts identified for retrieval purposes. Since we do not employ stemming or stopwording in our retrieval system, many of the synonyms are spelling variations of a particular concept. The queries were created with background knowledge, i.e. where necessary, *Annotator 1* looked up information about the event to determine a set of diverse terms. The created Indri queries are then deployed with the query likelihood retrieval model. Returned are the top $10,000$ documents (tweets) per query. In a post-processing step we filter out duplicates (tweets that are similar with cosine similarity > 0.9 to a tweet higher in the ranking) and then present the top 500 remaining tweets for judging to *Annotator 1* and *Annotator 2*. After

[2] TREC Microblog 2013 track: `https://github.com/lintool/twitter-tools/wiki/TREC-2013-Track-Guidelines`

[3] Wikipedia Current Events Portal, `http://en.wikipedia.org/wiki/Portal:Current_events`

[4] Indri is a query language supported by the Lemur Toolkit for Information Retrieval, `http://www.lemurproject.org/`.

the manual annotation process, the duplicates are injected into the relevance judgments again with the same relevance score and subtopic assignment as the original tweet.

Table 1. Examples of (i) news events selected, (ii) the corresponding Indri queries to generate a diverse result ranking for annotation purposes, (iii) the adhoc queries used in the retrieval experiments, and, (iv) examples of subtopics found during the annotation process (not all identified subtopics are shown).

News Event Topics	Manually created Indri queries	Adhoc queries	Identified Subtopics
Hillary Clinton steps down as United States Secretary of State	#combine(<#1(hillary clinton) #1(hilary clinton) #1(secretary clinton) #1(secretary of state)> <#1(steps down) #1(step down) leave leaves resignation resigns resign #1(stepping down) quit quits retire retires>)	*hillary clinton resign*	Clinton's successor what may be next for Clinton details of resignation Clinton's political positions
Syrian civil war	#combine(<syria syrian aleppo daraa damascus homs hama jasmin baniyas latakia talkalakh> <#1(civil war) war unrest uprising protest protests protestors demonstration demonstrators rebel rebels rebellion revolt revolts revolting resistance resisting resist clash clashes clashing escalation escalate escalated fight fights fighting battle battles offensive>)	*syria civil war*	casualties positions of foreign governments infighting among rebels
Boeing Dreamliner battery problems	#combine(<#1(Boeing Dreamliner) #1(boeing 787) #1(787 dreamliner)> <test tests testing tested check checks checked trial trials try> <battery batteries lithium-ion #1(lithium ion)>)	*dreamliner battery*	battery incidents cause of battery problems criticism Boeing tests

The annotators split the 50 topics among them and manually determined for each of the 500 tweets whether or not they belong to a particular subtopic (and which one). Thus, we did not attempt to identify subtopics beforehand, we created subtopics based on the top retrieved tweets. Tweets which were relevant to the overall topic, but did not discuss one or more subtopics were considered non-relevant. For example, for the topic *Hillary Clinton steps down as United States Secretary of State* we determined the first tweet to be relevant for subtopic *what may be next for Clinton*, while the second tweet is non-relevant as it only discusses the general topic, but no particular subtopic:

1. *Hillary Clinton transition leaves democrats waiting on 2016 decision. Hillary Clinton left the state department* $< URL >$.

2. *Clinton steps down as secretary of state. Outgoing us secretary of state Hillary Clinton says she is proud of* $< URL >$.

Thus, during the annotation process, we focused on the content of the tweet itself, we did not take externally linked Web pages in the relevance decision into account - we believe that this makes our corpus valuable over a longer period of time, as the content behind URLs may change frequently. This decision is in

contrast to the TREC 2011 Microblog track, where URLs in tweets were one of the most important indicators for a tweet's relevance [17].

We note, that defining such subtopics is a subjective process - different annotators are likely to derive different subtopics for the same topic. However, this is a problem which is inherent to all diversity corpora which were derived by human annotators. In order to show the annotator influence, in the experimental section, we not only report the results across all topics, but also on a per-annotator basis.

At the end of the annotation process, we had to drop three topics, as we were not able to identify a sufficient number of subtopics for them. An example of a dropped topic is *2012-13 UEFA Champions League*, which mostly resulted in tweets mentioning game dates but little else. Thus, overall, we have 47 topics with assigned subtopics that we can use for our diversity retrieval experiments.

4 Topic Analysis

In this section, we perform a first analysis of the 47 topics and their respective subtopics. Where applicable, we show the overall statistics across all topics, as well as across the topic partitions according to the two annotators.

The Topics and Subtopics In Table 2, we list the basic statistics over the number of subtopics identified, while Figure 1 shows concretely for each topic the number of subtopics. On average, we find 9 subtopics per topic. The large standard deviation indicates a strong variation between topics with respect to the number of subtopics (also evident in Figure 1). On a per annotator basis we also observe a difference in terms of created subtopics: *Annotator 1* has a considerably higher standard deviation than *Annotator 2*. This result confirms our earlier statement - subtopic annotation is a very subjective task.

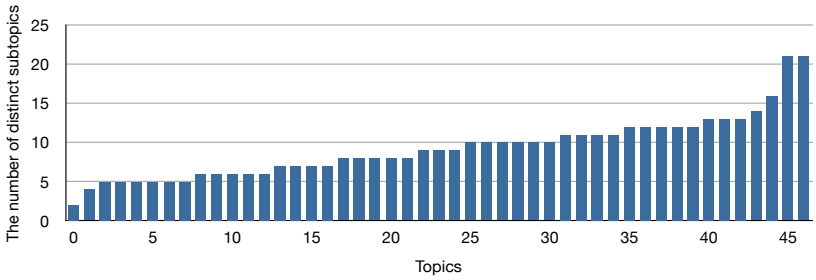
The topics yielding the fewest and most subtopics, respectively, are as follows:

- *Kim Jong-Un orders preparation for strategic rocket strikes on the US mainland* (2 subtopics)
- *Syrian civil war* (21 subtopics)
- *2013 North Korean nuclear test* (21 subtopics).

Table 2. Subtopic statistics.

	All topics	**Topics annotated by**	
		Annotator 1	***Annotator 2***
Av. num subtopics	9.27	8.59	9.88
Std. dev. subtopics	3.88	5.11	2.14
Min. num. subtopics	2	2	6
Max. num. subtopics	21	21	13

Fig. 1. Number of subtopics found for each topic.

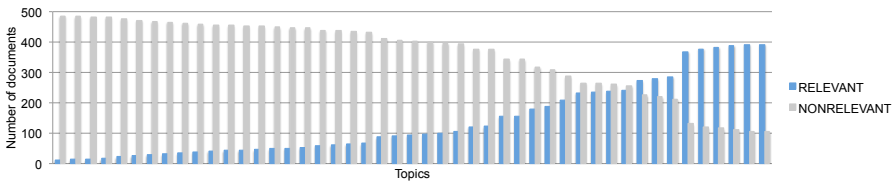


The annotators spent on average 6.6 seconds on each tweet in the annotation process and thus the total annotation effort amounted to 38 hours of annotations.

Apart from a very small number of tweets, each relevant tweet was assigned to exactly one subtopic - this is not surprising, considering the small size of the documents.

The Relevance Judgments In Figure 2 we present the distribution of relevant and non-relevant documents among the 500 tweets the annotators judged per topic[5]. Twenty-five of the topics have less than 100 relevant documents, while six topics resulted in more than 350 relevant documents. When considering the documents on the annotator-level, we see a clear difference between the annotators: *Annotator 1* judged on average 96 documents as relevant to a topic (and thus 404 documents as non-relevant), while *Annotator 2* judged on average 181 documents as relevant.

Fig. 2. Number of tweets per topic identified as (non-)relevant during the annotation process.

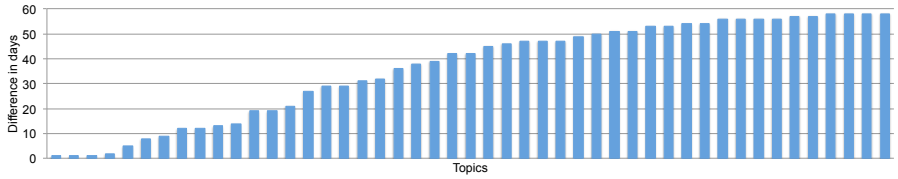


We also investigated the temporal distribution of the relevant tweets. In Figure 3 we plot for each topic the number of days that have passed between the first and the last relevant tweet in our data set. Since our data set spans a two-month period, we note that a number of topics are active the entire time (e.g. the topics *Northern Mali conflict* and *Syrian civil war*) while others are active froughly 24 hours (e.g. the topics *BBC Twitter account hacked* and *Eiffel Tower, evacuated due to bomb threat*). We thus have a number of short-term topics and a number of long-term topics in our data set. Inn contrast to the TREC Microblog

[5] As described earlier, the near-identical tweets that were removed to ease the annotation load are later added to the qrels again; they are not taken into account in the analysis presented here.

track 2011/12, we do not assign a particular querytime to each topic (therefore we implicitly assume that we query the data set one day after the last day of crawling). We do not consider this a limitation, as a considerable number of topics are covered across weeks.

Fig. 3. Difference in days between the earliest and the latest *relevant* tweet for each topic.



Diversity Difficulty Lastly, we consider the extent to which the search results can actually be diversified. Diversification does not only depend on the ambiguity or the underspecification of the query, it is also limited by the amount of diverse content *available in the corpus*. Golbus et al. [18] recently investigated this issue and proposed the *diversity difficulty* measure (*dd*) which is a function of two factors: the amount of diversity that a retrieval system can achieve at best and the ease with which a retrieval system can return a diversified result list. Intuitively, a topic has little inherent diversity if the maximum amount of diversity a retrieval system can achieve is small. A topic is considered "somewhat more diverse" by Golbus et al. in the case where a diverse result list can be achieved but it is difficult for the system to create one. A topic has a large amount of diversity if a retrieval system not tuned for diversity is able to return a diverse result list. These intuitions are formalized in a diversity formula with $dd \in [0, 1]$. A large score ($dd > 0.9$) indicates a diverse query, while a small score ($dd < 0.5$) either indicates a topic with few subtopics or a fair number of subtopics which are unlikely to be discovered by an untuned retrieval system. In Table 3 we present the diversity difficulty average and standard deviation our topics achieve - they are very similar for both annotators and also in line with the diversity difficulty scores of the TREC 2010 Web diversity track [18]. We thus conclude, that in terms of diversity our topic set presents a well constructed data source for diversity experiments.

Table 3. Diversity difficulty scores across all topics - a higher score is indicative of more diverse topics.

	All topics	**Topics assigned to**	
		Annotator 1	***Annotator 2***
Av. diversity difficulty	0.71	0.72	0.70
Std. dev. diversity difficulty	0.07	0.06	0.07

Finally, we observe that the diversity difficulty score of *long-term topics*, that is topics whose first and last relevant tweet cover at least a 50 day timespan, is higher ($dd_{long\text{-}term} = 0.73$), than the diversity difficulty score of *short-term* topics (the remaining topics) where $dd_{short\text{-}term} = 0.70$.

5 Diversification by De-Duplication

Having analyzed our corpus, we will now explore the diversification effectiveness of the recently proposed de-duplication framework for microblogs [7] on this data set.

5.1 Duplicate Detection Strategies on Twitter

In [7] it was found that about 20% of search results returned by a standard adhoc search system contain duplicate information. This finding motivated the development of a de-duplication approach which detects duplicates by employing (i) ***Sy***ntactical features, (ii) ***Se***mantic features, and (iii) ***Co***ntextual features in a machine learning framework[6]. By combining these feature sets in different ways, the framework supports mixed strategies named after the prefixes of the feature sets used: **Sy**, **SySe**, **SyCo** and **SySeCo**. Not surprisingly, the evaluation showed that the highest effectiveness was achieved when all features were combined.

Given an initial ranking of documents (tweets), each document starting at rank two is compared to all higher ranked documents. The duplicate detection framework is run for each document pair and if a duplicate is detected, the lower ranked document is filtered out from the result ranking.

5.2 Diversity Evaluation Measures

As researchers have been studying the diversification problem intensively on the Web, a number of measures have been proposed over the years to evaluate the success of IR systems in achieving diversity in search results. We evaluate our de-duplication experiments according to the following measures:

α-(n)DCG [14] This measure was adopted as the official diversity evaluation measure at TREC 2009 [1]. It is based on Normalized Discounted Cumulative Gain (nDCG) [19] and extends it by making the gain of each document dependent on the documents ranked above it.

Precision-IA [10] We evaluate the ratio of relevant documents for different subtopics within the top k items by the measure **Precision-IA**.

[6] The paper also consider the use of features derived from Web pages linked to in tweets. We ignore these features, as we did not consider URL content in the annotation process.

Subtopic-Recall [20] We report the subtopic recall (in short **S-Recall**) to show the number of subtopics covered by the top k documents. The measures ranges from 0 to 1, where larger values indicate a better coverage of subtopics.

Redundancy The measure shows the ratio of repeated subtopics among all relevant documents within the top k ranked documents. For diversity experiments, a lower redundancy value indicates a better performance.

5.3 Analysis of De-Duplication Strategies

We evaluate the different de-duplication strategies from two perspectives: (i) we compare their effectiveness on all 47 topics, and, (ii) we make side-by-side comparisons between two topic splits, according to the annotator and the temporal persistence. This enables us to investigate the annotator influence and the difference in diversity between long-term and short-term topics.

Apart from the de-duplication strategies we also employ three baselines: the **Automatic run** is a standard query likelihood based retrieval run (language modeling with Dirichlet smoothing, $\mu = 1000$) as implemented in the Lemur Toolkit for IR. The run **Filtered Auto** builds on the automatic run by greedily filtering out duplicates by comparing each document in the result list with all documents ranked above it - if it has a cosine similarity above 0.9 with any of the higher ranked documents, it is removed from the list. The de-duplication strategies also built on top of the Automatic Run by filtering out documents (though in a more advanced manner). All these runs take as input the adhoc queries (i.e. very short keyword queries) as defined in Table 1.

The only exception to this rule is the **Manual run** which is actually the run we derived from the manually created complex Indri queries that we used for annotation purposes with cosine-based filtering as defined above.

Overall comparison In Table 4 the results for the different strategies averaged over all 47 topics are shown. Underlined is the best performing run for each evaluation measure; statistically significant improvements over the *Filtered Auto* baseline are marked with † (paired t-test, two-sided, $\alpha = 0.05$). The *Manual Run* - as expected - in general yields the best results which are statistically significant in all measures at level @20.

We find that the de-duplication strategies *Sy* and *SyCo* in general outperform the baselines *Automatic Run* and *Filtered Auto*, though the improvements are not statistically significant. Not surpisingly, as the de-duplication strategies take *Automatic Run* as input, Preicision-IA degrades, especially for Precision-IA@20. On the other hand, in terms of lack of redundancy, the de-duplication strategies perform best. De-duplication strategies that exploit semantic features (*SySe* and *SySeCo*) show a degraded effectiveness, which is in stark contrast to the results reported in [7]. We speculate that the main reason for this observation is the recency of our corpus. Semantic features are derived from named entities (NE) recognized in the top-ranked tweets and queries. Since in [7] a corpus (documents and topics) from 2011 was used, it is likely that many more NEs were recognized (i.e. those NEs have entered the Linked Open Data cloud) than for our very

recent topics. As a concrete example, the topic *Syrian civil war* retrieves tweets which contain person names and locations important to the conflict, but they have not been added to standard semantics extraction services such as DBpedia Spolight[7].

Table 4. Comparison of different de-duplication strategies on our 47 diversity topics. Statistically significant improvements over the *Filtered Auto* baseline are marked with † (paired t-test, two-sided, $\alpha = 0.05$) for α-nDCG, Precision-IA and S-Recall. The Redundancy measure performs best when it is lowest.

Measure	α-nDCG		Precision-IA		S-Recall		Redundancy	
	@10	@20	@10	@20	@10	@20	@10	@20
Automatic Run	0.312	0.338	0.079	0.075	0.315	0.413	0.471	0.580
Filtered Auto	0.339	0.358	0.079	0.072	0.370	0.454	0.380	0.514
Sy	0.347	0.362	0.080	0.066	0.382	0.457	0.358	0.497
SySe	0.340	0.357	0.075	0.063	0.363	0.452	0.357	0.481
SyCo	0.346	0.360	0.080	0.065	0.381	0.464	0.371	0.478
SySeCo	0.341	0.358	0.077	0.064	0.365	0.457	0.376	0.489
Manual Run	0.386	0.443†	0.104†	0.099†	0.446	0.623†	0.482	0.601

Influence of Annotator Subjectivity and Temporal Persistence In Table 5, the results are shown when splitting the topic set according to the annotators. Here we find that although the absolute scores of the different evaluation measures for *Annotator 1* and *Annotator 2* are quite different, the general trend is the same for both. The absolute α-nDCG scores of the various de-duplication strategies are higher for *Annotator 2* than for *Annotator 1*, which can be explained by the fact that *Annotator 2*, on average, judged more documents to be relevant for a topic than *Annotator 1*. The opposite observation holds for the *Manual Run*, which can be explained by the inability of cosine filtering to reduce redundancy. Given that there are more relevant documents for *Annotator 2*'s topics, naturally the redundancy problem is more challenging than for *Annotator 1*'s topics.

Finally, Table 6 shows the results when comparing short-term and long-term queries. For long-term topics, the de-duplication strategies consistently outperform the baselines, while the same cannot be said about the short-term topics. We hypothesize that short-term topics do not yield a large variation in vocabulary (often a published news report is repeated in only slightly different terms) so that features which go beyond simple term matching do not yield significant benefits. Long-term topics on the other hand develop a richer vocabulary during the discourse (or the course of the event) and thus more complex syntactic features can actually help.

6 Conclusions

In this paper, we presented our efforts to create a microblog-based corpus for search result diversification experiments. A comprehensive analysis of the corpus showed its suitability for this purpose. The analysis of the annototators' influence on subtopic creation and relevance judgments revealed considerable subjectivity

[7] DBPedia Spotlight, `http://spotlight.dbpedia.org/demo/`

Table 5. Comparison of different de-duplication strategies when splitting the 47 topics according to the two annotators (due to the small topic size, significance tests were not performed).

Measure	α-nDCG		Precision-IA		S-Recall		Redundancy	
	@10	@20	@10	@20	@10	@20	@10	@20
Annotator 1								
Automatic Run	0.298	0.325	0.085	0.078	0.317	0.405	0.512	0.563
Filtered Auto	0.317	0.337	0.083	0.073	0.366	0.425	0.361	0.497
Sy	0.321	0.344	0.085	0.069	0.366	0.448	0.365	0.518
SySe	0.315	0.337	0.079	0.060	0.366	0.447	0.375	0.477
SyCo	0.318	0.346	0.086	0.067	0.359	0.466	0.339	0.464
SySeCo	0.321	0.344	0.083	0.062	0.358	0.466	0.362	0.460
Manual Run	0.442	0.489	0.127	0.111	0.537	0.667	0.451	0.582
Annotator 2								
Automatic Run	0.325	0.350	0.074	0.073	0.314	0.420	0.444	0.593
Filtered Auto	0.362	0.381	0.076	0.072	0.379	0.479	0.393	0.526
Sy	0.371	0.377	0.075	0.064	0.395	0.466	0.352	0.482
SySe	0.362	0.374	0.072	0.065	0.360	0.456	0.372	0.493
SyCo	0.371	0.373	0.075	0.063	0.400	0.462	0.369	0.482
SySeCo	0.359	0.371	0.073	0.066	0.371	0.448	0.386	0.509
Manual Run	0.338	0.403	0.087	0.090	0.367	0.583	0.505	0.615

Table 6. Comparison of different de-duplication strategies when splitting the 47 topics according to temporal persistence.

Measure	α-nDCG		Precision-IA		S-Recall		Redundancy	
	@10	@20	@10	@20	@10	@20	@10	@20
Long-term Topics								
Automatic Run	0.346	0.386	0.074	0.075	0.336	0.494	0.518	0.597
Filtered Auto	0.387	0.415	0.075	0.072	0.431	0.560	0.371	0.518
Sy	0.400	0.419	0.077	0.069	0.458	0.558	0.336	0.499
SySe	0.389	0.414	0.072	0.066	0.421	0.548	0.354	0.493
SyCo	0.401	0.416	0.078	0.068	0.459	0.554	0.358	0.486
SySeCo	0.386	0.412	0.074	0.069	0.417	0.545	0.376	0.501
Filtered Manual	0.373	0.431	0.084	0.087	0.416	0.596	0.457	0.619
Short-term Topics								
Automatic Run	0.293	0.311	0.082	0.075	0.304	0.367	0.437	0.571
Filtered Auto	0.312	0.326	0.081	0.072	0.336	0.393	0.402	0.510
Sy	0.318	0.329	0.081	0.065	0.338	0.400	0.388	0.495
SySe	0.312	0.325	0.077	0.061	0.330	0.397	0.375	0.464
SyCo	0.315	0.329	0.081	0.063	0.337	0.413	0.396	0.471
SySeCo	0.316	0.328	0.080	0.061	0.335	0.407	0.391	0.472
Manual Run	0.391	0.448	0.116	0.106	0.464	0.638	0.492	0.590

in the annotation process. At the same time though, the de-duplication retrieval experiments showed that the observed trends with respect to the different evaluation measures were largely independent of the specific annotator.

The performance of the de-duplication strategies and their comparison to the results reported in [7] indicate the importance of the feature suitability for the topic type (long-term vs. short-term topics and topic recency).

In future work we plan to further analyze the impact of the different strategies and the annotator subjectivity. We will also implement and evaluate the

de-duplication strategy with diversification approaches which have been shown to perform well in the Web search setting, e.g. [4, 5]. Furthermore, we will investigate the potential sources (influences and/or motivations) for the observed annotator differences.

References

1. Clarke, C.L.A., Craswell, N., Soboroff, I.: Overview of the trec 2009 web track. In: TREC '09. (2009)
2. Carterette, B., Chandar, P.: Probabilistic models of ranking novel documents for faceted topic retrieval. In: CIKM '09. (2009) 1287–1296
3. Slivkins, A., Radlinski, F., Gollapudi, S.: Learning optimally diverse rankings over large document collections. In: ICML '10. (2010) 983–990
4. Santos, R.L.T., Macdonald, C., Ounis, I.: Intent-aware search result diversification. In: SIGIR '11. (2011) 595–604
5. Santos, R.L.T., Macdonald, C., Ounis, I.: Aggregated search result diversification. In: ICTIR '11. (2011) 250–261
6. Teevan, J., Ramage, D., Morris, M.R.: #TwitterSearch: a comparison of microblog search and web search. In: WSDM '11. (2011) 35–44
7. Tao, K., Abel, F., Hauff, C., Houben, G.J., Gadiraju, U.: Groundhog day: Near-duplicate detection on twitter. In: WWW '13. (2013) 1273–1284
8. Cronen-Townsend, S., Croft, W.B.: Quantifying query ambiguity. In: HLT '02. (2002) 104–109
9. Bennett, P.N., Carterette, B., Chapelle, O., Joachims, T.: Beyond binary relevance: preferences, diversity, and set-level judgments. SIGIR Forum **42**(2) (2008) 53–58
10. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: WSDM '09. (2009) 5–14
11. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: SIGIR '98. (1998) 335–336
12. Zhai, C., Lafferty, J.: A risk minimization framework for information retrieval. Inf. Process. Manage. **42**(1) (2006) 31–55
13. Yue, Y., Joachims, T.: Predicting diverse subsets using structural svms. In: ICML '08. (2008) 1224–1231
14. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: SIGIR '08. (2008) 659–666
15. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: CIKM '09. (2009) 621–630
16. Clarke, C.L.A., Kolla, M., Vechtomova, O.: An effectiveness measure for ambiguous and underspecified queries. In Azzopardi, L., Kazai, G., Robertson, S.E., Rüger, S.M., Shokouhi, M., Song, D., Yilmaz, E., eds.: ICTIR '09. (2009) 188–199
17. Tao, K., Abel, F., Hauff, C., Houben, G.J.: What makes a tweet relevant for a topic? In: #MSM2012 Workshop. (2012) 49–56
18. Golbus, P., Aslam, J., Clarke, C.: Increasing evaluation sensitivity to diversity. Information Retrieval (2013) 1–26
19. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. **20**(4) (October 2002) 422–446
20. Zhai, C.X., Cohen, W.W., Lafferty, J.: Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In: SIGIR '03. (2003) 10–17