The Impact of Temporal Intent Variability on Diversity Evaluation

Ke Zhou[1], Stewart Whiting[1], Joemon M. Jose[1], and Mounia Lalmas[2]

[1] School of Computing Science, University of Glasgow, U.K.
[2] Yahoo! Lab. Barcelona, Spain
{zhouke,stewh,jj}@dcs.gla.ac.uk, mounia@acm.org

Abstract. To cope with the uncertainty involved with ambiguous or underspecified queries, search engines often diversify results to return documents that cover multiple interpretations, e.g. the car brand, animal or operating system for the query 'jaguar'. Current diversity evaluation measures take the popularity of the subtopics into account and aim to favour systems that promote most popular subtopics earliest in the result ranking. However, this subtopic popularity is assumed to be static over time. In this paper, we hypothesise that temporal subtopic popularity change is common for many topics and argue this characteristic should be considered when evaluating diversity. Firstly, to support our hypothesis we analyse temporal subtopic popularity changes for ambiguous queries through historic Wikipedia article viewing statistics. Further, by simulation, we demonstrate the impact of this temporal intent variability on diversity evaluation.

1 Introduction and Related Work

The uncertainty involved with ambiguous and underspecified queries is a common problem in information retrieval (IR) [1]. For example, a user specifying the ambiguous query topic 'presidents cup' may have an intent related to one of many possible subtopics: the President's Cup (or, synonymously Trophy) in golf, chess, tennis, hockey, lacrosse, football, and in many different countries (we analyse this example in Section 2). Without further clarification it is impossible to know the intent[1] of the user. To alleviate this problem, result diversification is a popular strategy used in IR to maximise for the effectiveness of the results for all users. When evaluating diversity, subtopic popularity and topical relevance of the documents are normally considered in conjunction. A system is rewarded if it provides relevant documents that cover the most popular subtopics as early as possible in the ranking. For example, for a ranking list of k documents, the intent-aware metric [1] (e.g. *nDCG-IA@k*) computes a traditional metric (e.g. *nDCG@k*) for each intent i (or subtopic) and then finally take an expectation based on the intent probabilities $P(i|q)$ (or, subtopic popularity).

$$nDCG\text{-}IA@k = \sum_i P(i|q) nDCG_i@k \tag{1}$$

However, current work generally evaluates diversity in a static environment, thereby making the assumption that subtopic popularity does not change over time.

[1] For consistency, we use *subtopic* to synonymously refer to an *intent* for the remainder of this paper.

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 821–824, 2013.

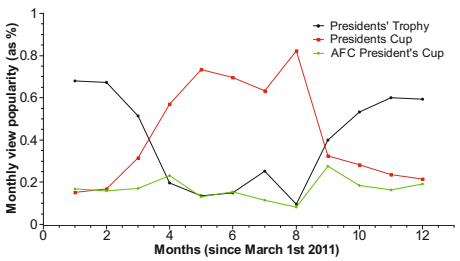

Fig. 1. Relative interest in '**Presidents Cup**' subtopics

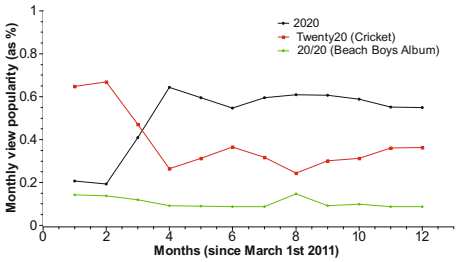

Fig. 2. Relative interest in '**2020**' subtopics

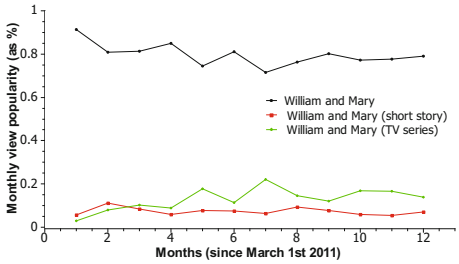

Fig. 3. Relative interest in '**William and Mary**' subtopics

Table 1. System ranking Spearmans' correlation, for topics of various temporal intent variability using *nDCG-IA*@10

temporal variability	*high*	*modest*	*low*
correlation	0.67 ▼	0.83 ▼	0.93

Time plays a central role in subtopic popularity for many topics [3]. For many query intents, temporal real-world events and phenomena such as news, politics and sport can have a major effect on what the user is most likely expecting. For instance, as shown in Figure 1, during the "President's Cup for golf", a user searching for the ambiguous topic would most likely be interested in the golf tournament. When the "President's Cup football competition" begins after the golf event, queries would be more likely related to the new event. As such, when evaluating a set of diversified systems, variance of subtopic popularity over time would likely affect system ranking. In this paper, we focus on ambiguous queries[2]. Our contributions are two-fold: **(1)** we analyse the *temporal intent variability* for ambiguous topics; **(2)** through a simulation we investigate the impact of temporality for *diversity evaluation*. To the best of our knowledge, this is the first work to quantitatively analyse the temporal variability of subtopic popularity in ambiguous queries using large-scale topic popularity data. Additionally, the impact of temporality on diversity evaluation measures has not been previously studied.

2 Temporal Intent Variability of Ambiguous Topics

The aim of this section is to quantify the temporal intent variability of ambiguous topics. Temporal intent variability in this paper refers to popularity changes between the

[2] We choose this for ease of implementation. The same techniques may be utilised to deal with multi-faceted queries.

subtopics of a single topic over time. For a whole user population, if most users' interests switch from one subtopic (e.g. "Presidents' Trophy for Golf") to the other (e.g. "Presidents' Cup Football Competition") after a short time period when issuing the same query/topic (e.g. "Presidents Cup"), we call this topic *highly variable*. Formally, given an ambiguous query q that contains a set of subtopics $S_q = \{s_1, s_2, ...s_n\}$, we want to quantify this variability $d(q,T)$ over a period T. We assume that T can be separated by a series of time intervals $T = \{t_1, t_2, ...t_m\}$. We also assume that we have the user view data $V_q^T = \{v_1^1, v_1^2, ...v_n^m\}$ where v_i^j is defined as the number of views for subtopic s_i at a given time period t_j. To quantify $d(q,T)$, for a given ambiguous topic we propose to track the change of the probability of interest of each subtopic over time. The mean of this change will then quantify the intensity of temporal intent variability for the given topic. Specifically, at time period t_j, we first calculate the probability of interest of a given subtopic s_i over all subtopics $P(s_i,t_j)$. Then we compute $d(q,T)$ as the mean of the standard deviation (denoted $stdev$) of each subtopic's $P(s_i,t_j)$ over T, as shown below:

$$P(s_i,t_j) = \frac{v_i^j}{\sum_{s_k} v_k^j} \qquad d(q,T) = \frac{1}{|S_q|} \sum_{s_i} stdev(s_i,T) \tag{2}$$

We adopt standard deviation for measuring the temporal probability change as it allows us to robustly measure the extent of the temporal deviation from the background level (i.e. mean) of a subtopic's interest.

Given the formal definition and approach, our implementation is as follows. First, we extract all the topics q and subtopics S_q from Wikipedia disambiguation pages. Second, we utilise the publicly available Wikipedia article user view data[3] to obtain V_q^T for each topic and corresponding subtopic. Specifically, we choose T within a one year span (March 2011 to March 2012), and separate T into a 52 week time-series. We use Wikipedia as the main resource as it provides substantial coverage of diverse topics and has been widely used in IR (e.g. intent prediction [2]). The hourly statistics of article page views provides us an alternative means to characterise the temporal aspects of topics when search query-log data is not available. Weekly aggregation of time provides appropriate granularity for both comprehensive (high recall) and robust (reduced noise) analysis. Since the unpopular topics and subtopics (those with very few user views) disproportionately affect relative measures of subtopic popularity, to remove noise we filter out the topics that received fewer than 10000 total subtopic views. Additionally we removed very unpopular subtopics, with a mean popularity percentage for the topic less than 5% over the one year span. Finally, we quantify all the topics by $d(q,T)$ as defined above and analyse the distribution of temporal intent variability.

Based on $d(q,T)$, we categorize all the ambiguous topics into three categories. Representative examples of *highly* ($d(q,T) > 0.15$), *modestly* ($0.05 < d(q,T) < 0.15$) and *lowly* ($0.0 < d(q,T) < 0.05$) temporally variable topics are shown in Figures 1, 2 and 3, respectively[4]. The number of topics of *high*, *modest* and *low* are 237 (1.41%), 5739 (34.0%) and 10887 (64.6%). Most of the *highly* variable topics are temporal topics where one or more subtopics are either part of, or themselves a major event during T.

[3] `http://dumps.wikimedia.org/other/pagecounts-raw/`

[4] We selected thresholds based on our observation of the data. More formal methods to temporally categorize the ambiguous topics are left for future work.

Figure 1 shows an example of this behaviour. Unlike *highly* variable topics, many *modestly* variable topics contain a single less pronounced event. As such, the popularity between subtopics varies much less over time. Figure 2 shows an example of this temporal phenomenon. For other topics of *low* variability, the subtopics' popularity remains comparatively static over time, as shown in Figure 3. Overall, these scenarios motivate us to investigate the impact temporal change on diversity evaluation.

3 Temporal Diversity Evaluation

Given various levels of temporal intent variability on ambiguous topics, we aim to investigate its impact on ranking diversified systems over time. We hypothesize that the more intense the temporal change is, the less correlated the system ranking over time will be. To study this, we follow the procedure as follows: **(1)**. we separate Wikipedia article user view data on a monthly basis within the year (March 2011 to March 2012) and we select topics and their corresponding subtopics' popularity for each month; **(2)**. for all the topics (100) from TREC web track 2009 and 2010, we simulate the subtopics' popularity for 12 months by assigning the subtopics' popularity from Wikipedia ambiguous topics; **(3)**. we randomly select 30 TREC participating systems; **(4)**. for each consecutive month pair (e.g. March-April), by utilizing *nDCG-IA*@10 as a diversity metric, we rank those systems based on different subtopic popularity over those two months and calculate their Spearman's correlation; **(5)**. We average all Spearman's correlations over a year and obtain the mean for all topics.

We select topics based on a given level of temporal variability and apply the above procedures to sets of topics with *high*, *modest* and *low* temporal intent variability. Significance (denoted by ▼) is computed using a paired t-test, with $p < 0.05$, with respect to results originated from the topic set of *low* temporal intent variability. The results are shown in Table 1. We can observe that: **(1)**. as what we expected, the correlation of system ranking is significantly lower on topic set of higher temporal intent variability; **(2)**. the correlation is not high, especially for topics of *high* temporal intent variability (0.67). This implies the need for development of time-aware diversity metrics.

4 Conclusions

This paper investigates the temporal intent variability of ambiguous queries, and its impact on diversity evaluation. We conclude that temporal subtopic popularity variability is modest or high for over 35% of ambiguous topics, and has considerably significant impact on diversity evaluation.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying Search Results. In: WSDM 2009, pp. 5–14 (2009)
2. Hu, J., Wang, G., Lochovsky, F., Sun, J.-T., Chen, Z.: Understanding Users Query Intent with Wikipedia. In: WWW 2009, pp. 471–480 (2009)
3. Whiting, S., Zhou, K., Jose, J., Alonso, O., Leelanupab, T.: CrowdTiles: Presenting Crowd-based Information for Event-driven Information Needs. In: CIKM 2012, pp. 2698–2700 (2012)