What makes a query temporally sensitive?

[Blind]

ABSTRACT

This work takes an in-depth look at the factors that affect manual classifications of "temporally sensitive" information needs. We use qualitative and quantitative techniques to analyze 660 topics from the Text Retrieval Conference (TREC) previously used in the experimental evaluation of temporal retrieval models. Regression analysis is used to identify factors in previous manual classifications. We explore potential problems with the previous classifications, proposing principles and guidelines for future work on temporal retrieval models.

1. INTRODUCTION

A growing body of information retrieval research argues that temporality should be modeled explicitly when scoring and ranking documents with respect to users' queries. Researchers have explored a variety of temporal retrieval models that explicitly incorporate time into document ranking [?, ?, ?]. They refer to general classes of "temporal queries" or "temporal information needs." Models have been proposed for "recency queries" [?, ?], "time-sensitive queries" [?], "implicitly temporal queries" [?], and "temporally biased queries" [?]. For evaluation, these studies rely on manual classifications of topics into temporal categories.

In this short paper, we take a deeper look into these manually classified topics to develop a clear understanding of *what makes a query temporally sensitive*? Previous manual classifications combine the temporal distribution of judged-relevant documents with common-sense notions of topic temporality without a clear explanation of the criteria or processes used in classification. If we cannot explain the processes being modeled, use of these manually classified topics for evaluation is of limited value.

To address this question, we analyze 660 topics from the Text Retrieval Conference (TREC) previously used in the experimental evaluation of temporal retrieval models. We employ qualitative and quantitative methods to identify topic characteristics that might affect the manual assessment of "temporal-sensitivity." The resulting coded topics are used in a set of regression analyses to assess the relationships between these characteristics and manually assigned categories. This paper's main contribution is an empirical assessment of the complexities that underpin temporal IR. This assessment helps us understand earlier temporal IR studies, while also suggesting novel ways to incorporate time effectively into retrieval.

2. TIME-SENSITIVE QUERIES

In this section, we review examples of studies focused on *temporal relevance*. The list of topics and collections used in each of the studies are listed in Table **??**.

Jones and Diaz [?] study the temporal characteristics of queries with the goal of query classification through the analysis of three TREC news collections and a web search engine log. They define three classes of queries based on the manual analysis of topics: temporally ambiguous (requesting multiple events), temporally unambiguous (requesting a single event), and atemporal (having no preference). Jones and Diaz manually classify 100 TREC topics based only on their title, description, and narrative fields. They also include 2003 Novelty track topics because they include topics classified as "event" or "opinion," which the authors suggest correspond to the "temporally unambiguous" and "atemporal" categories, respectively.

Dakka, Gravano, and Ipeirotis [?] investigate a broader class of queries which they refer to as "time-sensitive." They hypothesize that there are queries for which more relevant documents are found at specific points in time, not just recently. They manually examine the title, description and narrative of each topic and identify queries associated with specific news events. Only topics with > 20 judged-relevant documents are considered. If the topic information is insufficient to make a decision, they analyze the distribution of judged-relevant documents. The resulting classification consists of 86 temporally sensitive queries.

Efron and Golovchinsky [?] investigate additional models for recency queries. Topics are classified as "recency" if at least $2/3$ of the relevant documents occur after the median document time and the topic has a "bona fide temporal dimension" based on manual review, the specific criteria for which are not specified.

Finally, Peetz, Meij, and de Rijke [?] investigate the effect of temporal bursts in estimating query models. Building on the earlier studies, they evaluate their models using the previous manual classifications as well as a new collection based on TREC Blog06. As in the previous studies, the

SIGIR '2016 Pisa, Italy

authors construct a subset of "temporal" queries through manual evaluation of topic descriptions and relevant document distributions. No specific criteria for classification are given.

Topics	Collections	Studies
51-200	TREC Disks 1-2 AP (1988-89)	Jones & Diaz (2007)
301-450	TREC Disks 4-5 FT (1991-94); LA Times (1988-89)	Efron & Golovchinsky (2011); Dakka, Gravano & Ipeirotis (2012)
N1-100	AQUAINT Xinhua (1996-2000); NYT (1999-2000)	Jones & Diaz (2007)
851-1050	Blog06 (Dec 6, 2005 - Feb 21, 2006)	Peetz, Meij & de Rijke (2013)
MB1-110	Tweets 2011 (Jan 24, 2011 - Feb 8th, 2011)	Efron, Lin, de Vries (2014)

Table 1: TREC topics and Collections Used in Prior Temporal Retrieval Studies.

3. WHAT MAKES A QUERY TEMPORALLY SENSITIVE?

Given the complex landscape described in the previous section, what in general makes a query temporally sensitive? Dakka et al [?] present a compelling definition. A query is "time sensitive" if "the relevant documents for the query are not spread uniformly over time, but rather tend to be concentrated at restricted intervals." In other words, a query is temporally sensitive if relevant documents are more likely to occur at some points in time than others. This is an essential point, since many temporal retrieval models rely on the temporal distribution of results in document scoring. However, the distribution of relevant documents alone is not sufficient to determine true temporality (non-uniformity of document distributions with respect to time can be due to many factors). To address this, most of the studies listed above rely on common-sense notions of temporality based on the topic content considered independently of the distribution of relevant documents. A primary goal of the current study is to look deeper into these common-sense criteria with the aim of providing researchers a firmer basis for assessing which queries are likely to have a temporal relevance dimension.

4. METHODS

4.1 Qualitative coding

We use content analysis [?] to identify characteristics of TREC topics potentially associated with temporal sensitivity. 660 topics were selected from the TREC Ad-hoc, Novelty, Blog, and Microblog tracks, all previously used by researchers to evaluate temporal retrieval models. The complete set of topics used in this study are listed in Table **??** along with the temporal constraints of each collection or sub-collection.

Two of the authors participated in the development of the codebook and subsequent coding of topics. Codes were defined based on characteristics of topics expected to be related to temporal sensitivity, informed by the literature. During this process, code definitions were refined and clarified. In the final coding, only topic title and description were used. Of the 660 topics, 330 were coded by both coders to allow for inter-rater consistency analysis. The final codebook is too large to publish in this short paper, but is available online[1]. Coding was completed using the Dedoose[2] service. After coding all 660 topics, the topic/code matrix was exported for subsequent reliability and regression analysis, as described in the following sections.

An example of a coded topic from the 2004 Novelty test collection is presented in Figure **??**. This topic refers to a specific event and contains place entities as well as an explicit date. Topic N57 is categorized as an "event" by the TREC topic creator and is therefore an unambiguous temporal topic as defined by Jones and Diaz. In addition to coding the topics based on the defined codes, the coders assigned a temporal designation to the distribution of relevant documents for each topic. Non-parametric densities were fit to the temporal distribution of relevant documents for topics with more than 20 relevant documents, following Dakka et al. Each coder reviewed the relevant document distribution along with the total number of relevant documents for each topic and assigned one of four values based on subjective impressions about the degreee to which relevant documents were temporally constrained: too few observations (-1), low or no temporality (0), moderate temporality (1), and high temporality (2).

Title: $[(\text{East Timor})_{PlaceEntity}\text{Independence}]_{SpecificEvent}$

Description: $[(\text{East Timor})_{PlaceEntity}$ vote for independence from $(\text{Indonesia})_{PlaceName}$ in $(\text{August 1999})_{ExplicitDate}]_{SpecificEvent}$

Figure 1: TREC Novelty 2004 topic N57 example annotation

4.2 Reliability analysis

For this study, coder agreement is measured using Cohen's κ for the classification of the distribution of relevant documents. For the broader qualitative coding task, we use a variation of percent overlap, since coding is performed on arbitrary segments of text. We define *percent overlap* as:

$$overlap = \frac{m}{m + u_1 + u_2}$$

Where m is the number of excerpts assigned the same code by both coders, u_1 is the number of codes assigned to excerpts only by coder 1 and u_2 is the number of codes assigned to excerpts only by coder 2. If both coders assign no codes to a topic, it is considered perfect agreement. We report the macro (calculated over all topics) and micro (calculated as a per-topic average) overlaps. Per-code overlaps are used to characterize coder agreement within each code.

4.3 Relevant document distributions

In each of the four prior studies enumerated in Section 2, the authors acknowledge using the distribution of judged-relevant or pseudo-relevant documents in determining topic temporality. For this study, we use two different measures to analyze these distributions: the first-order time series autocorrelation (ACF) and the dominant power spectrum (DPS).

Jones and Diaz [?] use the ACF created by the temporal distribution of pseudo-relevant documents for a query as a predictor of query temporality. They note that queries with

[1] http://github.com/blind

[2] http://www.dedoose.com

strong inter-day dependencies will have high ACF values, indicating predictability in the time series.

Similarly, He, Chang, and Lim [?] use the DPS as a predictor of the "burstiness" of temporal features for event detection. The DPS is the highest power spectrum, estimated using the periodogram. The periodogram is the sequence of the squared magnitude of the Fourier coefficients $\|X_k\|^2$ indicating the signal power at frequency k/T in the spectrum.

In this study, both ACF and DPS measures are used to reduce the distribution of judged-relevant or pseudo-relevant documents to a single value for the regression analysis, as described in the next section.

4.4 Regression analysis

A primary goal of this study is to determine the characteristics that contribute to the manual judgment of topic temporality. We use logistic regression based on the generalized linear model (GLM) implementation in R. The predictors are binary presence indicators for each of the qualitative codes along with the ACF and DPS of the temporal distribution of true-relevant documents. The response variables are the binary temporal/non-temporal indicators manually assigned in the four studies. Model variables are selected using standard step-wise procedures based on the Akaike information criterion (AIC). Coefficients are reported using the log-odds and model fit is assessed using pseudo-R^2.

5. RESULTS

5.1 Codes

Our qualitative analysis suggests that three broad classes of features bear on query temporality: events, named entities, and explicit dates. It is intuitive that topics focused on specific and important events will have a higher degree of temporal relevance. Following the Topic Detection and Tracking definition, seminal events happen at specific times in specific places, often to individuals or other named entities (e.g., organizations). Perhaps the most essential code is the "SpecificEvent" – something important that happens at a particular time and place. Related to SpecificEvent is the "PeriodicEvent," which refers to an event that recurs periodically, such as the Super Bowl, World Cup, or Halloween. Jones and Diaz [?] note that many of the early adhoc queries were temporally ambiguous, referring to multiple events. We incorporate this concept through the "GenericEvent" code, which captures topics concerned with a class of specific events, such as earthquakes, elections, or strikes. While analyzing topics, it became apparent that some topics were likely to be inspired by a specific event, but without explicit reference in the topic description. This concept is captured through the "IndirectEventReference" code. The remaining codes are concerned with the identification of specific types of named entities, which are expected to have some association with topic temporality, and explicit dates.

5.2 Code distributions

Figure **??** summarizes the percent of topics in each test collection with each code assigned. We can see that the Novelty and Microblog collections have a higher percentage of specific events than the Blog and ad-hoc collections. The adhoc collections have a higher number of generic events, which supports the findings of Jones and Diaz [?]. The Blog, Novelty, and Microblog test collections each have larger numbers of named entities in the topic titles and descriptions.

Name	**Model**	R^2
Novelty	$-3.767 + 5.848 \cdot SpecEvt + 2.523 \cdot Other$	0.669
Novelty (Rel)	$-3.539 + 7.006 \cdot SpecEvt + 2.530 \cdot Other - 7.343 \cdot ACF$	0.706
Dakka	$0.134 + 0.878 \cdot Place$	0.019
Dakka (Rel)	$-0.917 + 0.393 \cdot DPS^{\blacktriangle}$	0.263
Efron	$-1.765 + 2.353 * Place^{\blacktriangle} + 1.410 \cdot Other^{\circ}$	0.181
Efron (Rel)	$-2.727 + 1.965 \cdot Place^{\blacktriangle} + 1.787 \cdot Other^{\vartriangle} + 0.163 \cdot DPS^{\blacktriangle}$	0.377
Peetz	$-0.336 + 1.682 * SpecEvt^{\circ} + 0.982 \cdot PerEvt + 0.672 \cdot Person - 0.6175 \cdot Org$	0.127
Peetz (Rel)	$-1.245 + 1.218 \cdot SpecEvt + 0.797 \cdot Period + 2.835 \cdot ACF^{\circ} + 0.002 \cdot DPS$	0.223

Table 2: Logistic regression models for each test collection without and with (Rel) ACF/DPS predictors. Model fit reported based pseudo-R^2 after stepwise variable selection based on AIC. Variable significance indicated by $p < 0.05(^{\circ}), < 0.01(^{\vartriangle}), < 0.001(^{\blacktriangle})$

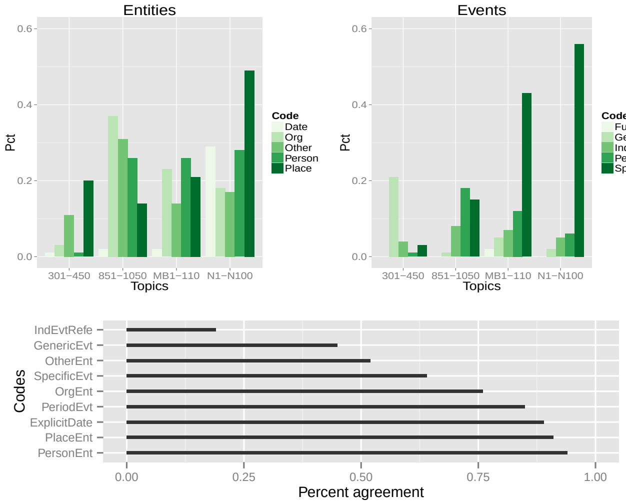

Figure 2: Percent of topics in each collection with codes assigned from the (a) entity code group and (b) events code group; (c) percent agreement by code.

5.3 Reliability

To assess coding reliability, a total of 1,244 codes were assigned to 330 topics by the two coders. Higher overlap indicates greater agreement between coders. The macro percent overlap is 0.71 and micro percent overlap is 0.83, indicating that overall our codes may be applied with good consistency. The per-code overlap is reported in Figure **??**(c). As expected, some codes have higher agreement than others. Specifically, personal names (0.94), locations (0.91), and explicit dates (0.89) have very high agreement whereas indirect event references (0.19) and generic events (0.45) have lower agreement.

5.4 Regression analysis

In this section, we report the results of the logistic regression analysis, predicting the manually assigned categories for each test collection. The resulting models are reported in Table **??**.

Collection	κ	ρ_{ACF}	ρ_{DPS}
AP	0.743	0.518	0.356
LA/FT	0.551	0.591	0.374
Blog	0.857	0.728	0.498
MB	0.806	0.692	0.354

Table 3: Cohen's κ for inter-coder agreement for classification of true-relevant document distributions. Pearson's ρ measuring correlation (average) between manual classifications and ACF/DPS values

For the 2003-2004 Novelty collection, the response variable is the manually assigned "opinion" (0) or "event" (1) categories. Following Jones and Diaz [?], we adopt "event" as the temporal category. Logistic regression analysis is performed with and without the ACF and DPS predictors. SpecificEvent and OtherEntity are significant predictors of the "event" category ($p < 0.01$), with a pseudo-R^2 of 0.669. Including the ACF of the true-relevant distribution is significant, with a minor improvement in model fit. The high pseudo-R^2 is unsurprising in this case, since the SpecificEvent code corresponds to the Novelty "event" category. It does, however, confirm our code definition.

Dakka et al manually classified "time-sensitive queries" for TREC topics 301-450. As reported in Table **??**, only the PlaceEntity code is a significant predictor of the manual classification. However, the pseudo-R^2 is very low (0.019). Dakka et al acknowledge examining the relevant document distributions for the LA Times and Financial Times sub collections. Including the DPS of the true-relevant document distribution increases the pseudo-R^2 to 0.263, suggesting that the relevant document distribution played a significant role in the manual classification.

Efron and Golovchinsky also classified topics 301-450, in this case focusing on the identification of "recency" queries. As reported in Table **??**, both PlaceEntity and OtherEntity are useful predictors of the temporal response. As with Dakka, including the DPS of the true-relevant distribution increases pseudo-R^2 from 0.181 to 0.377. This again suggests that the distribution of relevant documents played an important role in the determination of topic temporality.

Finally, we look at Peetz et al's classification of the Blog06-08 topics 850-1050. In this case, the SpecificEvent, PeriodicEvent, Person and Organization entities are useful predictors of the temporal category (pseudo-R^2=0.127). Including DPS improves model fit (pseudo-R^2=0.223), again suggesting that the distribution of relevant documents played a role in manual classification.

5.5 Relevant document distributions

As described in Section 3.1, non-parametric densities based on the temporal distribution of true-relevant documents are manually classified by two coders into four categories. The weighted Cohen's κ is calculated to assess agreement between the two coders. Average Pearson's correlation (ρ) measures the correlation between these manual classifications and the per-topic ACF/DPS values.

The results reported in Table **??** indicate moderate (0.40-0.60) to high (0.60-0.80) agreement between coders and higher correlation between the ACF and the manual classifications. These findings show that ACF and DPS do a good job representing the degree to which relevant documents are temporally constrained.

6. DISCUSSION AND CONCLUSIONS

In this study, we have tried to identify characteristics of TREC topics that can be used to predict manual classifications of "temporal sensitivity." Other researchers have classified topics without clear definitions or criteria. We have attempted to model these classifications by proposing features believed to indicate temporality. Features include the presence of different types of named entities, classes of events, and measures of the temporal distribution of judged relevant documents.

We were successful in modeling the "event" and "opinion" categories in the Novelty track, based primarily on our "SpecificEvent" code in the analyzed topics. Event codes were also found to be useful predictors of the classification of Peetz et al. In all cases, the distribution of relevant documents, represented by ACF or DPS, was consistently a significant predictor of topic temporality.

While these results are promising, we were unable to identify characteristics that fully explain the manual classification decisions. If we cannot explain the process that determines the classifications, it raises questions about the value of these test collections for evaluation. Specifically, how can we be clear that the queries previously identified as "temporally sensitive" are truly so? This ambiguity also limits the utility of previous temporal IR research, since it is unclear how to select queries for which the proposed models are well-suited.

7. ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation under Grant No. [blind]. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.