What makes a query temporally sensitive?

No Author Given

No Institute Given

Abstract. Keywords: We would like to encourage you to list your keywords within the abstract section

1 Introduction

A basic intuition in temporal information retrieval research is that time should be modeled explicitly when scoring and ranking documents with respect to users' queries. Users' criteria of recency, currency, and freshness have long been recognized as factors when judging relevance [2]. A number of studies have investigated the role of time in information retrieval using a variety of methods including query log analysis [9, 16, 12], temporal expression extraction [3, 8], temporal distributions of pseudo-relevant documents [7], and temporal retrieval models [11, 5, 4]. These researchers often refer to general classes of "temporal queries" and "temporal information needs." Models have been proposed and evaluated for "recency queries" [11, 5], "time-sensitive queries" [4], "implicitly temporal queries" [12], and "temporally biased queries" [7]. In a widely cited study, Jones and Diaz propose three different "temporal classes of queries" [7] including temporal, temporally ambiguous, and atemporal.

In this paper, we explore the question: *what makes a query temporally sensitive?* While many researchers have relied on the manual classification of topics, the methods used for classification are not clearly explained. Researchers refer to vague concepts of "newsworthiness" or the "*bona fide* temporal dimension" of topics without clearly explicating the criteria for classification.

To address this question, we analyze over 600 TREC topics previously used in the experimental evaluation of temporal retrieval models. We employ qualitative techniques to identify characteristics of topics that might affect manual assessment of "temporality." The resulting coded topics are used in a set of regression analyses to determine the specific relationships between these characteristics and manual assessment of topic temporality. Finally, we use the coded topics to predict when temporally-sensitive retrieval models might be effective.

This paper is structured as follows....

2 Temporal information needs

The concept of the *information need* has been extensively discussed in the information science literature and is widely used in the information retrieval community. While many theories have been proposed, no single definition is widely accepted. Synthesizing theories proposed by Taylor [?], Wilson [?], and Cole [?], here we define an *information need* as the unobservable motivation behind individual users' information seeking. Information needs reflect the individual user's current state of mind and context, including social and cultural environments. While information needs themselves are not observable, user's information seeking behaviors – including queries and document relevance judgements – are incomplete but observable evidence of the underlying need. Due to the individual nature of information needs, no two are identical, but given social and cultural contexts many needs are expressed by similar queries and satisfied by similar documents.

We believe that this definition captures the spirit of the concept of *information need* as commonly used in information retrieval research, particularly in the case of "topics" in the Text REtrieval Conference (TREC).

2.1 Time and relevance

There are numerous notions of temporality in information retrieval research, each of which requires different methodologies for analysis. In this study, we are primarily concerned with what we term as *temporal relevance*. We define this as the condition where information needs are satisfied by documents published at particular points in time. We distinguish this from *temporal topicality* which refers to information needs that are satisfied by documents about certain periods in time. Of course, an information need may combine the two conditions (e.g., find documents about the Wars of the Roses published in the 1800s). Examples of studies concerned with temporal topicality include [3, 8].

There are, of course, other notions of temporality in information retrieval research. For example, *temporal relevance dynamics* is concerned with how the relevance of a document changes with respect to an information need over time [13], *temporal query dynamics* is concerned with the study of how queries are issued over time [16, ?,14, 10]. Kulkarni et al [9] combine multiple sources of information – query logs, click-through logs, web crawler content changes, and human relevance judgments – to understand the relationship between query dynamics, relevance dynamics, and content dynamics.

2.2 Time-sensitive queries

Li and Croft's study of "recency queries" was the first in a line of research focused on what we have termed *temporal relevance*. Temporal relevance is concerned with conditions where documents published at particular points in time are considered more or less relevant than those published at other times. This condition is studied independent of or in conjunction with traditional topical retrieval models. In this section, we review how researchers in this area operationalize the concept of time and define the temporal characteristics of queries in their studies.

Li and Croft (2003) Li and Croft [11] hypothesize that some queries are "recency queries" where the most recently published documents are more likely to be relevant. They analyze queries associated with TREC topics 301-400 from TREC disks 4 and 5. Through the direct analysis of the temporal distribution of judged relevant documents, they classify 36 of the queries as recency queries because they have "more relevant documents in the recent past."

There are two problems with this approach. First, the distribution of judged relevant documents is not conclusive evidence that the query is temporally constrained. There are a variety of reasons why judged relevant documents may appear primarily in the recent past. Second, the authors did not recognized a potential problem in the underlying document collection. For their analysis, they used TREC 8 (disks 4 and 5), which consists of timestamped newswire documents from the Financial Times (1992-1994), Los Angeles Times (1989-1990), and the Foreign Broadcast Information Service (1994). Each of these sub-collections has distinct temporal characteristics that, when combined, may be misleading. Figure 1 shows the temporal distribution of results for query 301, a motivating example from their paper. Looking at the overall distribution of results from TREC 8, this appears to be a compelling example of a recency query. However, looking at the distribution of results for each sub-collection, we can clearly see that the "recent" results are those that appear in the FBIS collection. Further analysis indicates that the "recency" queries identified by Li & Croft are dominated by queries with results in FBIS, which has higher per-day document volumes than the other collections. This suggests that temporal retrieval models should be evaluated against individual sub-collections and that the temporal profile of collections is an important factor. Because of this, the results of Li & Croft's study as well as the identified "recency" queries should not be used further.

Jones and Diaz (2007) Jones and Diaz [7] 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 use the Associated Press (1988, 1989) and Wall Street Journal (1990-1992) collections from TREC disks 1 and 2 as well as the AQUAINT corpus with the 2003 Novelty track topics. They define three classes of queries based on the temporal distribution of pseudo-relevant documents: temporally ambiguous (requesting multiple events), temporally unambiguous (requesting a single event), and atemporal (having no preference). They employed annotators to manually classify 100 TREC queries based only on the topic title, description, and narrative. Specific criteria were not given. Interestingly, they found that all queries were only atemporal or temporally ambiguous. They incorporated the 2003 Novelty track because it includes topics classified as "event" or "opinion," which the authors found to correspond to the "temporally unambiguous" and "atemporal" categories.

We note a few things from this study, which will be discussed further later. First, the analysis is focused on news collections. Second, topics and relevance

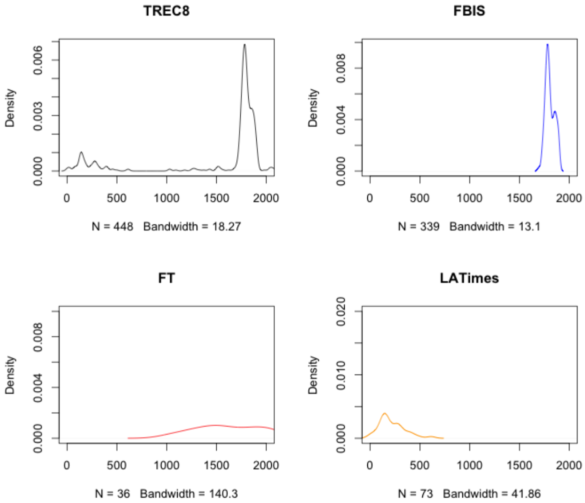


Fig. 1: Temporal distribution of results for topic 301 over TREC8 sub collections.

judgments were taken from existing TREC test collections. Third, the notion of "temporality" is somehow related to "events."

Dakka, Gravanos, and Ipeirtos (2012) Expanding on the work of Li and Croft, Dakka et al [4] 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 evaluate their models using a subset of TREC disks 4 and 5, manually identifying a subset of topics 301-450 that they consider to be "time-sensitive." To do so, they manually examine the title, description and narrative of each topic and identify queries associated with specific news events. If the topic information is insufficient to make a decision, they analyze the distribution of relevant documents. Only those queries with more than 20 matching documents[1] were considered. This resulted in a collection of 86 temporally sensitive queries. The document collection used for evaluation was restricted to the Financial Times (1991-1994) and Los Angeles Times (1989-1990) sub-collections, since they include timestamps.

Here we have another example of an analysis based on news collections where the central concept of a "time-sensitive" query is related to "specific news events." Queries were identified through manual analysis of the topic text and the ground truth relevance judgments, but no specific criteria aside from "newsworthiness" are given.

[1] based on conjunctive Boolean queries

Efron and Golovchinsky (2011) Also expanding on the work of Li and Croft, Efron and Golovchinsky [5] investigate additional models for recency queries. They use subsets of several TREC ad-hoc collections including the Associated Press documents from disks 1 and 2 with topics 101-200; Los Angeles Times and Financial Times documents from disks 4 and 5 with topics 301-450. They classify queries as "recency" or "non-recency" based on an analysis of the distribution of relevant documents. If at least 2/3 of relevant documents appear after the median document time, the query is considered a candidate for recency. Candidate queries are then manually reviewed to determine if they have a "bona fide" temporal dimension. However, the criteria for manual review were not specified. The authors developed a second test collection using the Twitter API. Two users of an experimental Twitter search engine were asked to create two types of queries: recency and non-temporal. Recency queries were defined as "queries where relevant tweets were necessarily written recently." Relevance judgments were collected via AMT.

In this case, we see a combination of news and social media where queries are classified based on manual review and analysis of judged-relevant document distributions.

Peetz, Meij, and Rijke (2013) In a more recent study, Peetz, Meij, and Rijke [15] investigate the effect of temporal bursts in estimating query models. Building on the above studies, they evaluate their models using the above test collections. In addition, they introduce the Blogs06 collection. As previously, the authors construct a subset of "temporal" queries through manual evaluation of topic descriptions and relevant document distributions. No specific criteria for classification are given.

Summary What can we conclude from these studies? First, each operationalizes time using the publication timestamp, which requires the availability of reliably timestamped document collections for evaluation. Most of these studies rely on collections of news articles, but more recent research also incorporates social media sources. Second, there is no standard approach for identifying "time-sensitive" queries. Some studies rely on the analysis of the temporal distributions of judged-relevant or pseudo-relevant documents, others on the manual analysis of topic descriptions, and some on a combination of the two. It is not clear from these studies how one determines whether a query is truly temporal or not. Third, temporally-sensitive queries are apparently related to "events." Jones and Diaz refer to the presence/absence of events in the manual analysis of topics. They also rely on the "event" category in the 2003 Novelty track. Dakka et al consider queries temporally sensitive if they are "associated with specific news events." Peetz, Meij, and Rijke assert that the proposed model "detects events." Unfortunately, none of these studies provides any definition or direction as to how to operationalize the concept of an "event."

The Topic Detection and Tracking (TDT) program was developed by NIST and ran for seven years. Two central tasks in the TDT program are 1) *topic* *detection* to detect emerging topics in news streams and 2) *topic tracking* to track those topics as they develop.

The 2004 TDT Annotation Manual provides the following definitions for topics and events:

1. *event*: a particular thing that happens at a specific time and place, along with all necessary preconditions and unavoidable consequences. A TDT event might be a particular plane crash, or a single meeting, or a particular court hearing.
2. *activity*: a connected set of events that have a common focus or purpose, happening at a specific place and time; for instance, a campaign, or an investigation, or a disaster relief effort.
3. *topic*: an event or activity, along with all directly related events and activities.

In TDT, events must be one of thirteen *seminal events* and specific guidelines are provided as to what types of events are considered related. Examples include elections, scandals, criminal cases, disasters, accidents, discoveries, and acts of violence. These TDT definitions inform the codebook development process used in this study.

3 What makes a query temporally sensitive?

We return now to the motivating question: what makes a query temporally sensitive? Dakka et al provide 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 one point in time than another. However, there are various reasons that relevant documents might not be uniformly distributed, which might not be related to true temporality. For example, the constraints of the document collection might affect relevant document distribution. Dakka et al refer to “newsworthiness” in their classification of topics. Efron and Golovchinsky refer to a “bona fide temporal dimension.” Of course, all of these ignore the case of temporal query dynamics, where a query may also be temporally sensitive depending on when it is issued.

We propose a broader definition: a query is “temporally sensitive” if the documents that are relevant to that query depend on time. This can be because the documents are about a particular period in time or because the document is published at a particular time. A query can also be temporally sensitive because of when it was issued.

In the previous sections, we review how researchers approach investigations of the role of time in information retrieval. A few common characteristics emerge:

1. When was the query issued?
2. What are the temporal constraints of the collection and sub-collections?

3. Does the query contain an explicit or implicit temporal expression?
4. Is the query focused on a 1) a specific event, 2) a set of events, 3) a type of event, or 4) no events?
5. Is the event a seminal or singular event?
6. Is the query focused on a specific entity?
7. Are documents at some points in time more likely to be relevant than documents at other points in time?

In the next sections we report the results of an analysis of over 600 TREC topics used in temporal information retrieval research.

4 Methods

In the studies reviewed above, researchers rely on existing test collections, such as those available through TREC, to evaluate temporal retrieval models. In each study, topics are manually categorized as temporal or non-temporal to assess model performance. The purpose of this study is to further investigate the characteristics of topics deemed temporal. To achieve this, we use a combination of qualitative content analysis and regression analysis, as described below.

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. These topics were selected because they have been used previously by researchers in temporal retrieval research and have associated manual classifications. The complete set of topics are listed in Table 1.

Topics	**Collections**
51-200	TREC Disks 1-2 AP 88-89; WSJ 87-92
301-450	TREC Disks 4-5 FT 91-94; LA Times 88-89
N1-100	AQUAINT Xinhua 1996-2000; NYT 1999-2000; AP 1999-2000
851-1050	Blog06
MB1-110	Tweets 2011

Table 1: TREC topics used in this study

Two of the authors participated in the development of the codebook and subsequent coding of the topics. Codebook development began with a preliminary reading of all topic titles, descriptions and narratives. Codes were defined based on characteristics of topics expected to be related to topic temporality. Of the 660 topics, 330 were coded by both coders. During this process, code definitions were refined and clarified. In the final coding, only topic title and description were used. The final codebook is presented in Table 4 in the appendix. Coding was completed using the Dedoose[2] service. Following coding, the topic/code matrix was exported for subsequent reliability and regression analysis.

4.2 Reliability analysis

Coding reliability is measured using a variation of percent overlap. In this study, conventional measures such as Cohen's κ or Krippendorf's α are not applicable, as the coding is performed on arbitrary segments of text in each topic. We define the *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, this is considered perfect agreement. We report the macro overlap calculated over all topics, the micro overlap calculated as a per-topic average, and per-code overlaps to understand coder agreement within each category.

4.3 Relevant document distributions

In each of the four prior studies, authors acknowledge using the distribution of judged-relevant documents in determining topic temporality. For this study, we use two different measures to represent this distribution: time series autocorrelation and the dominant power spectrum.

Jones and Diaz [7] use the first-order autocorrelation (ACF) of the time series created by the temporal distribution of pseudo relevant documents for a query as a predictor of query temporality. They note that queries with strong inter-day dependencies will have high ACF values, indicating predictability in the time series.

He, Chang, and Lim [**?**] use the power spectrum of the dominant period of a time series (DPS) as a predictor of the "burstiness" of temporal features. The DPS is the highest power spectrum, estimated using the periodogram.

In this study, both of these measures are used to represent the distribution of judged-relevant documents in a regression analysis, described in the next section.

4.4 Regression analysis

For this study, a logistic regression is performed for each test collection using the generalized linear model (glm) implementation in R. The predictors are binary presence indicators for each code from the topic/code matrix along with the ACF and DPS values. The response variable is the binary temporal/non-temporal indicator manually assigned in the previous studies. Model variables

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

are selected using standard step-wise procedures. Predictors are reported using the standard log-odds. Model fit is assessed using 10-fold cross validation and reported using prediction error.

4.5 Predicting temporal model effectiveness

In the first part of the study, we investigate whether the coding strategy can be used to predict the manual classification of topics. In this part, we develop similar to models to predict whether to use a temporal retrieval model for each topic. We use standard query likelihood [**?**] and the kernel density estimate KDE temporal model [6] to determine query temporality. If the average precision (AP) of the KDE model score is greater than the AP of the standard query likelihood score, topics are classified as "temporal." If the QL model is more effective, the topic is classified as "non-temporal." As in the previous section, logistic regression analysis is used. The ACF and DPS of the pseudo-relevant document distribution are used to approximate the relevant document distribution.

5 Results

What are the characteristics of topics that affect the manual assessment of topic temporality?

5.1 Codes

Our qualitative analysis suggests three broad classes: events, named entities, and explicit dates. A basic intuition is that topics focused on a specific and important events will have a higher degree of temporal relevance. Following the TDT 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 code, which refers to an event that recurs periodically, such as the Super Bowl, World Cup, or Olympics. Jones and Diaz [7] noted that many of the early ad-hoc 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 the event is not referenced 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.

5.2 Code distributions

Table **??** summarizes the percent of topics in each test collection with each code assigned. From these results, we can see that the Novelty and Microblog test collections have a higher percentage of specific events than the Blog and ad-hoc collections. The ad-hoc collections have a higher number of generic events, which supports the findings of Jones and Diaz [7]. The Blog, Novelty, and Microblog test collections each have larger numbers of named entities in the topic titles and descriptions.

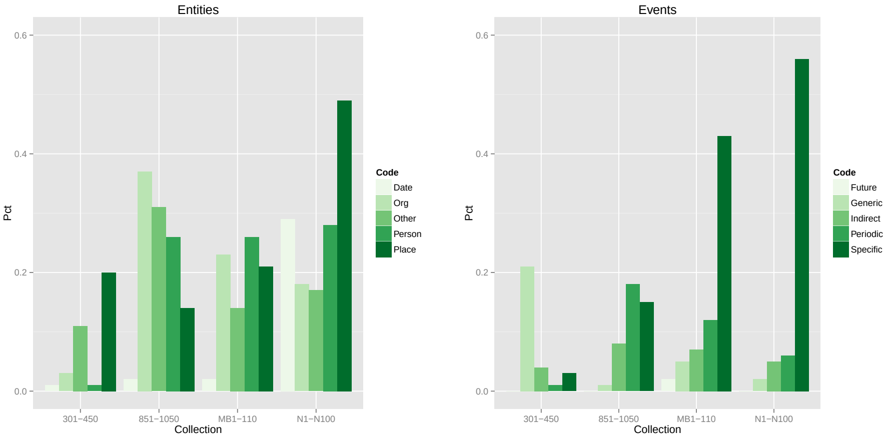


Fig. 2: Percent of topics in each collection with codes assigned from the (left) entity code group and (right) events code group

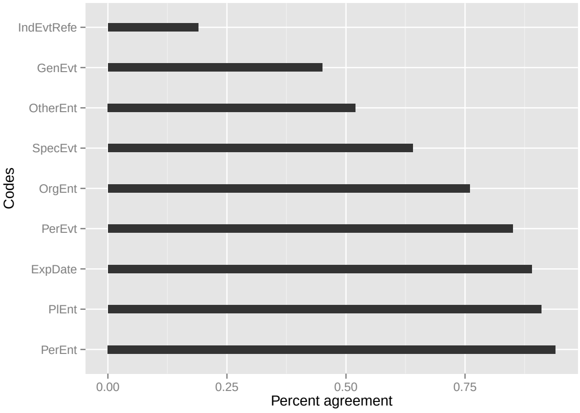
5.3 Reliability

To assess reliability, a total of 1244 codes were assigned to 330 topics by two coders. The macro percent overlap is 0.71 and micro percent overlap is 0.83. The per-code overlap is reported in Table **??**. Higher overlap indicates greater agreement between coders. 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, were report the results of a logistic regression analysis, predicting the manually assigned categories for each test collection. The resulting models are reported in Table **??**.

AP all 0.2166 0.2171 0.0005 0.2183 t = -5.4211, df = 149, p-value = 1.167e-07 LATimes all 0.1976 0.1984 0.0008 0.2003 t = -2.0653, df = 149, p-value = 0.02032 Novelty 03 all 0.2947 0.2853 -0.0094 0.2986 t = -3.3995, df = 49, p-value



Name	**Model**	**AIC**	**Pred. Erro**
Novelty	-3.767 + 5.848*SpecEvt + 2.523*OtherEnt	52	0.070
Novelty (ACF/DPS)	-3.539 + 7.006*SpecEvt + 2.530*OtherEnt - 7.343*ACF	49	0.070
Dakka	0.134 + 0.878*PlaceEnt	205	0.427
Dakka (ACF/DPS)	-0.917 + 0.393*DPS	155	0.233
Efron	-1.765 + 2.353*PlaceEnt + 1.410*OtherEnt	150	0.207
Efron (ACF/DPS)	-2.727 + 1.965*PlaceEnt + 1.787*OtherEnt + 0.163*DPS	118	0.160
Peetz	-0.336 + 1.682*SpecEvt + 0.982*PerEvt + 0.672*PerEnt -0.6175*OrgEnt	192	0.327
Peetz (ACF/DPS)	-1.245 + 1.218*SpecEvt + 0.797*PerEvt + 2.835*ACF + 0.002*DPS	171	0.313

Table 2: Logistic regression models for each test collection with and without ACF/DPS variables. Model fit reported based on AIC, cross-validation prediction error, and pseudo-R^2.

Collection	**Model**	**AIC**	**Pred. Error**	R^2
Tweets 2011-12	0.722 - 2.332*GenEvt$^{\uparrow}$	141	0.318	0.043
Novelty 2003-4	-2.658 + 0.920*PlaceEntity$^{\uparrow}$ + 1.183*SpecificEvent$^{\uparrow}$ + 2.677*ACF$^{\uparrow}$	130	0.38	0.10
AP	-0.767 + 0.040*DPS$^{\uparrow}$	204	0.43	0.04
Blog06	-1.047 + 0.848*PeriodicEvent + 0.017*DPS$^{\uparrow\uparrow}$	195	0.35	0.065
LATimes	0.689 + 3.431*ExplicitDate - 0.711*PlaceEntity - 0.116*DPS$^{\uparrow}$	199	0.41	0.054

Table 3: Logistic regression models predicting KL/KDE for each test collection. Model fit reported based on AIC, cross-validation prediction error, and pseudo-R^2. $^{\uparrow}$ indicates p ¡ 0.05. $^{\uparrow\uparrow}$ indicates p ¡ 0.01

= 0.0006751 Novelty 04 all 0.3689 0.3643 -0.0046 0.3728 t = -3.9596, df = 49, p-value = 0.0001213 Blog06 all 0.2978 0.2975 -0.0003 0.3013 t = -3.6519, df = 149, p-value = 0.0001799 Tweets 2011 all 0.2251 0.2418 0.0167 0.2448 t = -4.7387, df = 48, p-value = 9.756e-06 Tweets 2012 all 0.1794 0.1907 0.0113 0.1953 t = -4.9802, df = 58, p-value = 3.018e-06

Novelty We begin with the 2003-2004 Novelty topics. In this case, the response variable is the manually assigned “opinion” (0) or “event” (1) categories. Following Jones and Dias [7], we adopt “event” as the temporal category. Logistic regression analysis is performed with and without the ACF and DPS variables. Looking at Table **??**, SpecificEvent is a useful predictor of the “event” category (p ¡ 0.01). This is unsurprising, since the definition of the SpecificEvent code corresponds to the Novelty “event” category. Including ACF has a minimal effect.

Dakka et al Dakka et al provided manual classification of “time-sensitive queries” in TREC topics 301-450. As reported in Table **??**, the DPS variable has a considerable effect in predicting the temporal category. No other variables are significant in this model. This suggests that either 1) Dakka et al relied heavily on the distribution of relevant documents in determining the classification or 2) our model is missing an important explanatory component.

Efron and Golovchinsky Efron and Golovchinsky also classified topics 301-450, in this case identifying only “recency” queries. As reported in Table **??**, both the PlaceEntity and OtherEntity codes are useful predictors of the temporal category. Inclusion of the DPS variable substantially improves model fit. This suggests that the distribution of relevant documents played some role in the determination of topic classes.

Peetz et al Finally, we look at Peetz et al’s classification of the Blog06-08 topics 850-1050. In this case, the SpecificEvent and PeriodicEvent code are useful predictors of the temporal category. Including the ACF and DPS variables improves model fit, again suggesting that the distribution of relevant documents played a role in manual classification.

5.5 Predicting when to use temporal retrieval models

In the previous section, we report the effectiveness of using the coded topics to predict manually classified temporal categories. In this section, we develop similar to models to predict whether to use a temporal retrieval model for each topic. We use standard query likelihood [**?**] and the kernel density estimate KDE temporal model [6] to determine query temporality. If the average precision (AP) of the KDE model score is greater than the AP of the standard query likelihood score, topics are classified as "temporal." If the QL model is more effective, the topic is classified as "non-temporal." As in the previous section

Instead of testing these retrieval models in conventional evaluation, we instead use the resulting classification for comparison to the manual classifications and the results of the above described content analysis.

The content analysis relies on a manual review and interpretation of the topic text with some investigation of related contexts. The original manual classifications relied on a combination of manual interpretation of topic text and interpretation or heuristic classification based on the distribution of true-relevant documents. The retrieval models rely on the distribution of pseudo-relevant documents. As a final measure, we adopt the cross-correlation function (CCF) used by Amodeo et al [1] to measure the correlation of the true-relevant and pseudo-relevant document distributions.

The final analysis will include a regression analysis of the above features including the qualitative codes, previously assigned classifications, automatic classification based on retrieval performance and the CCF.

6 Conclusions

7 Acknowledgments

This section is optional

References

1. G. Amodeo, G. Amati, and G. Gambosi. On relevance, time and query expansion. *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, page 1973, 2011.
2. C. Barry and L. Schamber. Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing & Management*, 34(2):219–236, 1998.
3. K. Berberich, S. Bedathur, O. Alonso, and G. Weikum. A Language Modeling Approach for Temporal Information Needs. In *Proceedings of the 32Nd European Conference on Advances in Information Retrieval*, pages 13–25, 2010.
4. W. Dakka, L. Gravano, and P. Ipeirotis. Answering General Time-Sensitive Queries. *IEEE Transactions on Knowledge and Data Engineering*, 24(2):220–235, Feb. 2012.
5. M. Efron and G. Golovchinsky. Estimation methods for ranking recent information. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*, page 495, 2011.
6. M. Efron, J. Lin, J. He, and A. D. Vries. Temporal Feedback for Tweet Search with Non-Parametric Density Estimation. *umiacs.umd.edu*, pages 33–42, 2014.
7. R. Jones and F. Diaz. Temporal profiles of queries, July 2007.
8. N. Kanhabua and K. Nø rvåg. A comparison of time-aware ranking methods. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*, page 1257, 2011.

9. A. Kulkarni, J. Teevan, K. M. Svore, and S. T. Dumais. Understanding temporal query dynamics. *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*, page 167, 2011.
10. V. Lavrenko, M. Schmill, and D. Lawrie. Mining of concurrent text and time series. *. . . on Text Mining*, pages 2–9, 2000.
11. X. Li and W. B. Croft. Time-based language models. *Proceedings of the twelfth international conference on Information and knowledge management - CIKM '03*, page 469, 2003.
12. D. Metzler, R. Jones, F. Peng, and R. Zhang. Improving search relevance for implicitly temporal queries. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, (1):700, 2009.
13. S. Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.
14. N. Parikh and N. Sundaresan. Scalable and near real-time burst detection from eCommerce queries. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, page 972, 2008.
15. M. Peetz and M. D. Rijke. Cognitive temporal document priors. *Advances in Information Retrieval*, 2013.
16. M. Shokouhi. Detecting Seasonal Queries by Time-Series Analysis Categories and Subject Descriptors. In *SIGIR 11*, pages 1171–1172, 2011.

Appendix: Codebook

Code	**Description**	**Examples**
Specific Event	Something significant that happens at a specific time and place. Code title and description in concert, even if title does not contain event specifics.	Mount Pinatubo erruption on June 15, 1991; 2008 State of the Union; Hurricane Hugo
Generic Event	Use this code when the topic refers to more than one specific event or a class or type of event. Only use this code if every instance of the event type would be newsworthy (i.e., a specific event) and the central topic of a news article.	Earthquakes, volcano erruptions, elections, disputes, strikes
Indirect Event Reference	Apply this code to indicate when the topic might be indirectly referring to a *specific* event. Use only if you need to turn to external information to identify potential specific events (e.g., your personal knowledge, wikipedia). Do not use if specific event information is contained in the description.	Legally assisted suicide, related to Kevorkian controversy. Partial birth abortion ban, related to partial birth abortion ban legislation. Surrogacy related to Baby M.
Periodic Event	Apply this code to indicate when an event is periodic, recurring at regular, predictable intervals. Never double-code as SpecificEvent or as an entity even though periodic events are often named entities.	Super bowl, Nobel awards, Oscars, State of the Union
FutureEvent	Apply this code to indicate when a topic refers to a future predicted specific event. Never double-code as SpecificEvent.	2020 Fifa, 2016 Summer Olympics
Person Entity	Apply this code to identify personal names in topics.	President Bush; Sasha Cohen;
Place Entity	Apply this code to identify places in topics. Limit to proper names. Also apply to references to nations, governments or government bodies.	Peru; Africa; African; European; Atlanta
Organization Entity	Apply this code to identify organizations in topics. Limit to proper names. Do not apply to references to governments (e.g., United States), use PlaceEntity instead.	Hitachi Data Systems; U.S. Congress;
Other Entity	Apply this code to named entities that are not people, places or organizations. Limit to proper names. Includes movies, books., etc.	Hubble Telescope; The Avengers; Euro
Explicit Date	Apply this code to identify explicit dates.	1988; June 15, 1991; October 2007; Monday

Table 4: Codebook