What makes a query temporally sensitive?

ABSTRACT

Here is the abstract

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Keywords

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 [11, 18, 14], temporal expression extraction [3, 10], temporal distribution of pseudo-relevant documents [9], and temporal retrieval models [13, 7, 6]. These researchers refer to general classes of "temporal queries" and "temporal information needs." Models have been proposed and evaluated for "recency queries" [13, 7], "time-sensitive queries" [6], "implicitly temporal queries" [14], and "temporally biased queries" [9]. In a widely cited study, Jones and Diaz propose three different "temporal classes of queries" [9].

In this paper, we explore the question what *what makes a query temporally sensitive?*

To address this question, we analyze over 600 topics previously used in the experimental evaluation of temporal retrieval models. In four previously published studies [9, 7, 6, 17], researchers manually classified standard TREC topics as either "temporal" or "non-temporal." In this study, we employ qualitative techniques to identify and annotate characteristics of topics that might affect manual assessment of "temporality." The resulting coded topics are used in a regression analysis to determine the specific relationships between these characteristics and manual assessment of topic temporality. Finally, we use the coded topics to predict which topics might benefit from temporally-sensitive retrieval models.

This paper is structured as follows. In section 2, we review the concept of information needs. In section 3, we review the role of time in information retrieval research.

To address this question, we first review how researchers have approached investigations of the role of time in information retrieval. We survey the literature, focusing on methods of analysis, definitions, and operationalizations of time-related concepts. We then report the results of our analysis of over 600 topics previously used in experimental evaluation.

2. ON INFORMATION NEEDS

The concept of *information needs* 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).

3. A BRIEF HISTORY OF TIME IN IR

Early evidence of temporal information needs can be found in descriptive and subject cataloging practice employed in library catalogs. Library cataloging predates online information and in many ways reflects the historical needs of users. Descriptive and subject cataloging practice capture a variety of temporal characteristics of published resources including publication dates, copyright dates as well as chronological terms intended to facilitate subject searching based on particular periods of time. While publication dates cap-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM '2016 Indianapolis, IN USA

Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

ture when a resource was published or created, chronological terms and subdivisions (e.g., Wars of the Roses, 1455-1485) in subject headings indicate if a resource is *about* a particular period in time.

With regard to time, descriptive cataloging is primarily concerned with production dates. This includes publication dates, distribution dates, and copyright dates. In the event that publication dates are not present, catalogers often provide approximate dates based on their knowledge and experience. Publication date information allows users to search for resources based on temporal constraints (e.g., date range) or to rank results based on the most recent or oldest publication dates.

Subject cataloging, on the other hand, is concerned with the topical content of the work. Subject headings lists, such as the Library of Congress Subject Headings (LCSH), include a variety of different methods for specifying temporal content [**?**]. Subject headings allow users to search for resources based on temporal (or chronological) content.

These cataloging practices suggest two distinct classes of temporal information needs: those based on the ability to find resources published at particular points in time and those based on the ability to find resources about particular periods in time. Of course, an information need may combine the two (e.g., find books about the Wars of the Roses published in the 1800s). We will refer to these as *temporal relevance* and *temporal topicality.*

3.1 Time and relevance

In two separate studies of user criteria for judging document relevance, Barry and Schamber [2] both found *recency* (or *currency*) to be an important characteristic for some users and information needs. They define this as "the extent to which users judged information to be current, recent, up-to-date, or timely." [2] In Barry's analysis of 18 interview respondents, "currency" was the sixth most-mentioned out of 21 different relevance criteria. In Schamber's study of 30 respondents, it was the second most-mentioned to of 10 different relevance criteria.

The concept of *recency* is related to the later concept of *freshness* found in research focused on web search [5]. In web search, *freshness* is related to the time when a web page or link was last updated. [4] Research has consistently found that users prefer (and thus judge as more relevant) resources that are more frequently and recently updated. Prior to the web, *recency* was primarily represented by document publication dates.

Another role of time in the judgment of relevance is suggested by Mizzaro [15], who observes that, as the user learns things or the problem changes, a document may not be relevant to a query at one point in time, but be relevant later (or vice versa). This type of user-dependent temporal relevance can be described as *temporal relevance dynamics* – how the relevance of a document with respect to an information need may change over time. While Mizarro is concerned mainly with changes in the users' state of mind, as we will discuss later, what is considered relevant can also change at an aggregate level for many users.

So far, we have identified three broad classes of temporal information needs: 1) instances where document publication dates matter; 2) instances where the topics are periods in time; and 3) instances where what is relevant changes over time. Following Jones and Diaz, we can add to this the negative case 4) when time is not a factor at all.

3.2 Temporal query dynamics

Informally, the study of *dynamics* is concerned with characteristics that stimulate changes in a system or process. We use *temporal dynamics* to refer to the specific study of change with respect to time. Whereas *temporal relevance dynamics* focuses on the study of how what is relevant to users changes over time, *temporal query dynamics* focuses on the study of queries over time. [11]

Studies of temporal query dynamics generally focus on the analysis of query streams in the form of search engine logs. These studies have identified classes of temporal patterns in query streams (e.g., seasonal/periodic [18], bursts [19], trends [16]) as well as methods for identifying these patterns and using them for the classification of queries or correlation with external events [12, 19].

One of the more compelling studies in this area is that of Kulkarni et al [11]. The authors 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. They study how time impacts which queries are issued, which documents are deemed relevant, and how changes in the content of documents relate to relevance judgments.

Queries are observable evidence of users' information needs. Patterns of similar queries and similar judged-relevant documents over time suggest that certain queries reflect common underlying information needs. These studies indicate that some queries are repeated over time, others occur at particular points in time, and that query intents also change over time. That is, the same query surface form may represent distinct information needs and be satisfied with different documents at different points in time.

Returning to our motivating question, these studies suggest additional conditions. The "intent" of commonly-issued queries, and therefore which documents satisfy these queries, can change over time. Changes in query intent can be one-time, as the primary or dominant intent changes due to shifts in popularity, or periodic: at different times of day, days of the week, or days in the year. In each of these cases, documents that are likely to be relevant to a particular query depend on when the query is issued.

3.3 Temporal topicality

Metzler et al [14] investigate *implicitly year qualified queries* which they indicate are a subset of "temporal queries." They define a "year-qualified query" as a query containing a year and an "implicitly year-qualified query" as a query where the user might have a specific year in mind. Examples include "miss universe," "olympics," "easter," and the names of annual conferences. The study is based on the analysis of query logs, with time operationalized by the presence or absence of years in the query string.

Year qualified queries are in one sense temporally-topical queries. Users are primarily concerned with content about the year qualified topic, not with information published at a particular point in time. In the examples provided, year qualified queries are also periodic or seasonal, referring to events that occur on an annual basis. It is reasonable to suggest that implicit year qualified queries are a type of recency query, where users are interested in the most recent

or up-to-date information about a periodic event. This suggests that queries concerning periodic events may combine temporal-topicality and temporal relevance.

Taking a different approach, Berberich et al [3] propose an extension to the language modeling framework for information retrieval that incorporates information derived from temporal expressions. Example queries include "fifa world cup 1990s" or "13th century crusades." Their basic approach is to extract temporal expressions from queries and documents and convert them to the time domain. Documents are then scored based on their query likelihood with regard to the textual content combined with the likelihood given the temporal coverage of the document in the time domain.

The authors develop a test collection based on the New York Times Annotated Corpus. Two different human intelligence tasks (HITS) were developed using the Amazon Mechanical Turk (AMT) service. In the first task, users were given entities related to a subset of topics (sports, culture, technology, and world affairs) and asked to specify a temporal expression for the entity. In the second task, users were shown a temporal expression and asked to select and entity from the topics. From these tasks, 40 queries were selected. Examples include "boston red sox [october 27, 2004]," "pearl harbor [december, 1941]," "sewing machine [1850s]," and "muhammed [7th century]." Relevance assessments were collected using AMT.

The test collection developed by Berberich et al is important, as it is used for this and several subsequent studies. The authors rejected the idea of using existing test collections because they need "a specific class of information needs" represented by queries with explicit temporal expressions. There are several interesting characteristics of this test collection. First, the document collection is a corpus of news articles. As we will see in the next section, news articles are a common document type for temporal retrieval research. Second, the queries are developed around "entities": people, organizations, places, and products. A review of the 40 queries indicates that over 50% contain either names of people or organizations. This suggests one of three things: 1) queries in general are primarily about named entities; 2) explicit temporal expressions occur primarily with respect to named entities; or 3) it's easier to construct queries with temporal expressions with respect to named entities.

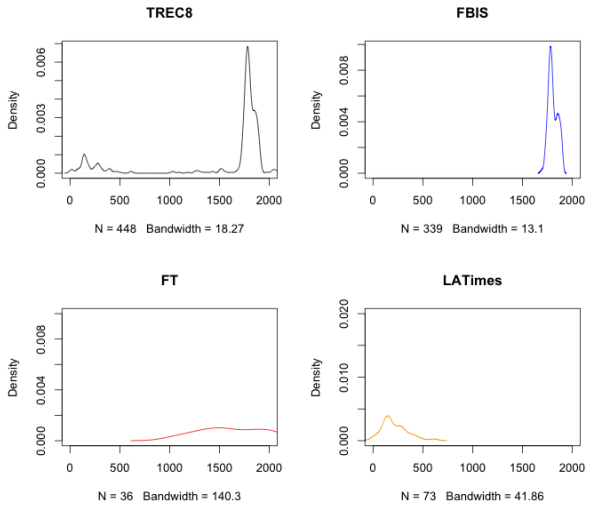
Kanhabua and Norvag [10] explore "time-aware retrieval models" using Berberich's NYT corpus. They use a similar justification, that TREC "queries are not time-related, and judgments are not targeted towards temporal information needs." In addition to temporal expressions found in the document content, they also incorporate temporal information based on the document publication time. Both content time and publication time are used with respect to the explicit temporal expression found in the query.

3.4 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 operationalize the concept of time and define the temporal characteristics of queries in their studies.

Li and Croft [13] 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 X 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 subcollection, 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. 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 [9] 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 defined three classes of queries based on the temporal distribution of pseudo-relevant documents: temporally ambiguous (requesting multiple events), tempo-

rally unambiguous (requesting a single event), and atemporal (had no preference). They employed annotators to manually classify 100 TREC queries based only on the topic title, description, and narrative. Interestingly, they found that the queries were only atemporal or temporally ambiguous. The 2003 Novelty track 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 and search engine logs. Second, topics and relevance judgments were taken from existing TREC test collections. Third, the notion of "temporality" is somehow related to "events."

Expanding on the work of Li and Croft, Dakka et al [6] investigate a broader class of queries which they refer to as "time-sensitive." They hypothesize that there are queries where more relevant documents are found at specific points in time, not just recent. They evaluate their models using two different test collections: a subset of TREC disks 4 and 5 and a custom test collection based on the Newsblaster news service. Similar to Li and Croft, for the TREC collection they manually identify a subset of queries from 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. The second test collection was developed based on a six-year archive of Newsblaster, covering news articles from September 2001 to December 2006. The authors recruited five journalists who volunteered queries. From 125 collected queries, 76 were identified as time-sensitive. Relevance judgments were collected using AMT.

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.

Further expanding on the work of Li and Croft, Efron and Golovchinsky [7] 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.

In a more recent study, Peetz, Meij, and Rijke [17] investigate the effect of temporal bursts in estimating query models. Building on the above studies, they evaluate their models using several previously used collections (AP, LA Times, Financial Times). 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.

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 sources. Of the 10 studies reviewed, 8 use newswire collections, 3 use Twitter, 2 use blog collections, and 1 uses a standard web collection[2]. 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."

3.5 Topic detection and tracking

Much the research discussed so far is concerned primarily with using temporal characteristics in ad-hoc retrieval. A related area of research also concerned with time is that of document filtering. Whereas ad-hoc retrieval focuses on ranked lists of documents searched retrospectively, filtering is focused on decisions about document relevance made at specific points in time. Given a time-ordered stream of documents, systems must determine whether to emit or not emit a document when it is received, under certain constraints. Sub-areas of filtering research are often focused on specific tasks such as topic or event detection.

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. TDT is the precursor to the recent Knowledge Base Acceleration (KBA) and Temporal Summarization tracks in TREC. In all of these cases, the goal is to monitor a stream of documents for changes that occur to specific information needs or topics.

The 2004 TDT Annotation Manual provides the following

[1] based on conjunctive Boolean queries

[2] This is by no means an exhaustive list of work in this research area

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. In this sense, the *topic* is the seminal event. The seminal event types and examples from the 2004 Annotation Manual are listed in Table 5.

The TDT research program is concerned with real-time filtering of time-ordered streams of documents, In this case, there is no topic or information need. Instead, systems are challenged to identify and track topics over time. Topics are related to the concept of an *event* or *seminal event*. The collection is composed of timestamped new articles.

4. WHAT MAKES A QUERY TEMPORALLY SENSITIVE?

Queries represent a user's underlying information need and are issued over a collection at a particular point in time. A single information need is often represented by multiple related queries issued over time. In this sense, a set of queries can be seen as evolving over the course of a single user's session as they learn more or the problem changes. Users – and as a result their information needs – also change over time. As a result, users refine queries as well as the criteria for what constitutes relevance.

Queries reflect broader social contexts. Many users may issue the same or similar queries and be satisfied with similar documents. For common queries, we can consider the behavior of an "average user." While information needs are necessarily individual, the success of information retrieval models depends on common patterns and the behavior of average users. We can study the same query issued by multiple users to identify patterns in documents judged relevant to that query.

The documents that satisfy the average user for a particular query may change over time. Consider the example of the query "u.s. open." The intent of the query changes depending on the year: golf in June, tennis in September. Consider the example of the query "flawless," which might refer to the film before and the song after November 2013. In this case, the dominant sense changes at a particular point in time. By examining only the surface form of the query, this means that the documents considered relevant to the query "flawless" change over time.

There is an assumption in both of these examples that the document collection is also changing – that more documents published after November 2013 refer to the song. The dominant sense of a particular query is reflected not only in the query surface form, but also in the content the user is searching. Queries and documents both reflect and are motivated by events in the outside world.

We can come up with examples of queries with changes in the dominant sense over static collections. Consider users searching a collection of recipes. Certain recipes may be more or less popular, and therefore relevant to the query, depending on the time of year. For instance, cold recipes preferred in summer and warm recipes in winter, or seasonal recipes around the holidays. In this case, the query intent changes because of the time the query was issued.

Collections are composed of documents, all published at particular points in time. Collections therefore have temporal characteristics such as start dates, end dates, and distributions of documents over time (e.g., number of documents published per day). Some collections are static while others change as new documents are added. In other cases, including the web, the documents themselves change.

So, let's return 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. This ignores 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 be temporally sensitive because of when it was issued.

In the previous section, 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 3) 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 800 TREC topics used in temporal information retrieval research.

5. 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

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

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.

5.1 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.

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 6 in the appendix. Coding was completed using the Dedoose [3] service. Following coding, the topic/code matrix was exported for subsequent reliability and regression analysis.

5.2 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.

5.3 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.

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

Jones and Diaz [9] 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 interday 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.

5.4 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 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.

5.5 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 [8] 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.

6. RESULTS

6.1 Code distributions

Table 2 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 [9]. The Blog, Novelty, and Microblog test collections each have larger numbers of named entities in the topic titles and descriptions.

6.2 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 3. 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

Topics	ExpDate	OrgEnt	OtherEnt	PersonEnt	PlaceEnt	FutureEvt	GenericEvt	IndEvtRef	PerEvt	SpecEvt
301-450	0.01	0.03	0.11	0.01	0.20	0.00	0.21	0.04	0.01	0.03
851-1050	0.02	0.37	0.31	0.26	0.14	0.00	0.01	0.08	0.18	0.15
N1-N100	0.29	0.18	0.17	0.28	0.49	0.00	0.02	0.05	0.06	0.56
MB1-110	0.02	0.23	0.14	0.26	0.21	0.02	0.05	0.07	0.12	0.43

Table 2: Percent of topics with each code assigned by topic group

Code	Overlap
PersonEntity	0.94
PlaceEntity	0.91
ExplicitDate	0.89
PeriodicEvent	0.85
OrganizationEntity	0.76
SpecificEvent	0.64
OtherEntity	0.52
GenericEvent	0.45
IndirectEventReference	0.19

Table 3: Per-code percent overlap

high agreement whereas indirect event references (0.19) and generic events (0.45) have lower agreement.

6.3 Regression analysis

6.3.1 Novelty

Novelty 2003-2004: In this analysis, manually assigned codes are used to predict the Novelty "event" and "opinion" categories.

Without ACF/DPS:

```
              Estimate Std. Error  Pr(>|z|)
(Intercept)     -3.854      1.036  0.000199 ***
ExplicitDate     2.170      1.211  0.073268 .
OtherEntity      2.097      1.360  0.123093
SpecificEvent    5.290      1.115  2.1e-06 ***

AIC 49.287
CV prediction error 0.070
```

With ACF/DPS:

```
              Estimate Std. Error Pr(>|z|)
(Intercept)     -3.747      1.086  0.000557 ***
ExplicitDate     2.326      1.202  0.052889 .
OtherEntity      2.278      1.297  0.079082 .
SpecificEvent    6.606      1.444  4.78e-06 ***
ACF             -8.113      3.804  0.032918 *

AIC: 45.708
CV prediction error 0.080
```

6.3.2 Efron and Golovchinsky

Efron and Golovchinsky (51-200, FT)

Without ACF/DPS:

```
                  Estimate Std. Error Pr(>|z|)
(Intercept)        -1.8372     0.2806  5.88e-11 ***
OrganizationEntity  2.0011     1.3462  0.1371
OtherEntity         1.5807     0.6609  0.0168 *
PlaceEntity         2.2546     0.4861  3.52e-06 ***
IndirectEventRef    1.6820     1.2703  0.1855
PeriodicEvent     -17.5539  1542.5833  0.9909

AIC: 149.45
CV prediction error: 0.213
```

With ACF/DPS:

```
                        Estimate Std. Error Pr(>|z|)
(Intercept)          -3.01773     0.45589  3.61e-11 ***
OrganizationEntity    2.74178     1.71794  0.110496
OtherEntity           2.21050     0.76411  0.003817 **
PlaceEntity           2.15790     0.63735  0.000710 ***
IndirectEventRef      3.19011     1.64329  0.052222 .
PeriodicEvent       -20.44363 1438.13865  0.988658
SpecificEvent        -2.80536     1.87915  0.135467
DPS                   0.19832     0.05409  0.000246 ***

AIC: 115.05
CV prediction error 0.167
```

6.3.3 Dakka et al

Without ACF/DPS: Regression analysis indicates no features are useful predictors of manually assigned categories.

With ACF/DPS:

```
                Estimate Std. Error Pr(>|z|)
(Intercept)     -0.94500    0.26426  0.000349 ***
SpecificEvent   16.68523 1516.27888  0.991220
DPS              0.37944    0.08651  1.15e-05 ***

AIC: 153.04
CV prediction error: 0.24
```

6.3.4 Peetz et al

Without ACF/DPS:

```
                   Estimate Std. Error Pr(>|z|)
(Intercept)          -0.3958     0.2834  0.1625
ExplicitDate         16.2401  1083.7318  0.9880
OrganizationEntity   -0.6614     0.3929  0.0923 .
PersonEntity          0.6829     0.4355  0.1169
GenericEvent         16.6400  1665.1334  0.9920
PeriodicEvent         1.0686     0.5139  0.0376 *
SpecificEvent         1.5576     0.6818  0.0223 *

AIC: 189.07
CV prediction error: 0.313
```

With ACF/DPS:

```
                Estimate Std. Error Pr(>|z|)
(Intercept)   -1.296e+00  3.185e-01  4.72e-05 ***
ExplicitDate   1.581e+01  1.161e+03  0.9891
GenericEvent   1.648e+01  1.542e+03  0.9915
PeriodicEvent  8.655e-01  5.441e-01  0.1117
SpecificEvent  1.193e+00  7.589e-01  0.1161
ACF            2.758e+00  1.441e+00  0.0556 .
DPS            2.168e-03  1.216e-03  0.0745 .
```

```
AIC: 170.72
CV prediction error: 0.30
```

6.4 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 [8] 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.

7. EXAMPLES

[Note: I don't expect to include all of this in a submitted paper. For now, it's helping me to work through key examples of concepts that inform the study.]

In this section, we review selected TREC topics from these collections and compare classifications from four different studies. Jones and Diaz (51-200), Efron and Golovchinsky (100-200, 301-450), Dakka et al (301-450), Peetz et al (850-950, 1001-1050).

> Topic: 57
> Title: MCI
> Description: Document will discuss how MCI has been doing since the Bell System breakup.

Topic 57 is focused on a named entity, MCI, in relation to a specific event, the Bell System breakup, which occurred in January 1984. The temporal sensitivity of this topic depends on two things: the temporal constraints of MCI and the temporal constraints of the underlying collections. This topic was created in 1992 for TREC1, which includes AP 88-89 and WSJ 87-92. MCI Communications was founded in 1963. Given that the seminal event is well outside the temporal constraints of the collections, there is no reason to believe that relevant documents would occur at one time period more than another. Jones and Diaz classify this query as 'atemporal.'

> Topic: 139
> Title: Iran's Islamic Revolution - Domestic and Foreign Social Consequences
> Description: Document will report on the religious, legal, cultural, and social consequences of Iran's Islamic Revolution within Iran and abroad.

Topic 139 is focused on the consequences of a specific event. Iran's Islamic Revolution occurred in 1979, long before the temporal constraints of the test collection. The concept of 'consequences' is vague and might refer to a variety of events or activities. There is no indication from the title or description that this topic is concerned with a particular point in time. Efron and Golovchinsky classify this as a recency query.

> Topic: 301
> Title: International Organized Crime
> Description: Identify organizations that participate in international criminal activity, the activity, and, if possible, collaborating organizations and the countries involved.

Topic 301 is the motivating example form Li and Croft discussed above. The topic description makes no reference to specific events, only a broad class of events (criminal activity). There is no indication that relevant documents would occur at one point in time more than another. Efron and Golovchinsky classify this as a non-recency query, Dakka et al classify it as a time-sensitive query.

> Topic: 374
> Title: nobel prize winners
> Description: Identify and provide background information on Nobel prize winners.

Topic 374 is concerned with a periodic event (Nobel prize). The Nobel prizes are announced in October with a ceremony and banquet on December 10. While the topic is concerned with a list of winners, it's reasonable to expect that more relevant documents would be found around the time of the announcement or banquet. The topic was developed in 1998 for a test collection consisting of LA Times (88-89) and FT (91-94). Efron and Golovchinsky classify this as non-recency. Dakka et al classify it as time-sensitive.

> Topic: 409
> legal, Pan Am, 103
> What legal actions have resulted from the destruction of Pan Am Flight 103 over Lockerbie, Scotland, on December 21, 1988?

Topic 409 is concerned with a set of general events (legal actions) in relation to a specific event, the Lockerbie bombing. The topic was created in 1999 and the test collection consists of Los Angeles Times (88-89) and Financial Times (91-94). Legal actions develop over longer periods of time and it is unclear whether relevant documents would be more likely to appear at one point or another. If anything, this topic may be temporally sensitive with respect to the LA Times collection, but not the Financial Times collection. Efron and Golovchinsky classify this as a recency query. Dakka et al classify it as a time-sensitive query.

Topic: 410
Schengen agreement
Who is involved in the Schengen agreement to eliminate border controls in Western Europe and what do they hope to accomplish?

Topic 410 is concerned with a specific event, the Schengen agreement, which was signed in 1985. The agreement was supplemented in 1990 and implemented in 1995. The query was developed in 1999 and the test collections consist of LA Times (88-89) and FT (91-94). There is perhaps some reason to expect that relevant documents are more likely to occur around the time of the 1990 agreement in the FT collection. Efron and Golovchinsky classify this as a recency query. Dakka et al classify it as a time-sensitive query.

Topic: N16
Title: Kenya Tanzania Embassy bombings
Type: event
Description: U.S. Embassy bombings in Africa, 1998

Jones and Diaz accept all Novelty "event" topics as temporally unambiguous. This topic refers to two specific bombings that occurred in Africa in 1998. The events are specific and the topic description contains an explicit date. There is reason to believe that, in the AQUAINT corpus (1996-2000), relevant documents would be more likely to appear near or after the time of the bombings.

Topic: N19
Title: Elian Gonzalez Cuba
Type: opinion
Description: What are opinions about returning Elian Gonzalez to Cuba?

Jones and Diaz classify all Novelty "opinion" topics as temporally ambiguous. In this case, the topic refers to a specific event – the Elian Gonzalez affair. Gonzalez arrived in November 1999 and was returned to Cuba on June 28, 2000. There is reason to believe that relevant documents in the AQUAINT corpus (1996-2000) would be more likely to appear near near his return in June.

Title: larry summers Description: Find opinions on Harvard President Larry Summers' comments on gender differences in aptitude for mathematics and science.

Peetz et al classify this query as non-temporal. Larry Summers made controversial comments at an economics conference in January 2005, which is outside the constraints of the Blog06 collection (12/05 - 2/06). It is possible that the topic would be again discussed one year later, on the anniversary of the event.

Topic: 853
Title: state of the union
Description: Find opinions on President Bush's 2006 State of the Union address.

Peetz et al classify this query as temporal. The State of the Union address occurred on January 31, 2006. It is a specific, seminal event that occurred within the constraints of the Blog06 collection.

So, what can we learned from this small sample of topics? To determine the temporal sensitivity of a query we need to understand:

1. when the query was issued
2. temporal constraints of the collection(s)
3. temporal constraints of central entities or events
4. whether events are seminal or periodic
5. whether the query is concerned with a seminal event or a class of events

8. CONCLUSIONS

9. ACKNOWLEDGMENTS

This section is optional

10. 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] K. Berberich, M. Vazirgiannis, and G. Weikum. Time-Aware Authority Ranking. *Internet Mathematics*, 2(3):301–332, Jan. 2005.

[5] N. Dai and B. D. Davison. Freshness Matters : In Flowers , Food , and Web Authority. pages 114–121.

[6] 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.

[7] 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.

[8] 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.

[9] R. Jones and F. Diaz. Temporal profiles of queries, July 2007.

[10] 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.

Collection	Topics	Future	Recent	Opinion	Specific	Periodic	Type	Explicit	Implicit	NE	Total
Ad-hoc	51-200, 251-450	5	16	8	20	3	129	10	2	121	350
Novelty	1-100	0	0	38	55	4	14	27	3	76	100
Blog06	851-950	1	1	100	14	21	6	2	0	86	100
Microblog	1-225	1	0	0	111	26	23	3	1	165	225

Table 4: Results

[11] 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.

[12] V. Lavrenko, M. Schmill, and D. Lawrie. Mining of concurrent text and time series. *... on Text Mining*, pages 2–9, 2000.

[13] 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.

[14] 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.

[15] S. Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.

[16] 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.

[17] M. Peetz and M. D. Rijke. Cognitive temporal document priors. *Advances in Information Retrieval*, 2013.

[18] M. Shokouhi. Detecting Seasonal Queries by Time-Series Analysis Categories and Subject Descriptors. In *SIGIR 11*, pages 1171–1172, 2011.

[19] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos. Identifying similarities, periodicities and bursts for online search queries. *Proceedings of the 2004 ACM SIGMOD international conference on Management of data - SIGMOD '04*, page 131, 2004.

APPENDIX

Seminal event	Examples
elections	a specific political campaign, election day coverage, inauguration, voter turnouts, election results, protests, reaction
scandals/hearings	media coverage of a particular scandal or hearing, evidence gathering, investigations, legal proceedings, hearings, public opinion coverage
legal/criminal cases	the crime itself, arrests, investigations, legal proceedings, verdicts and sentencing
natural disasters	weather events (El Nino, tornadoes, hurricanes, floods, droughts), other natural events like volcanic eruptions, wildfires, famines and the like, rescue efforts, coverage of economic or human impact of the disaster
accidents	transportation disasters, building fires, explosions and the like
acts of violence or war	a specific act of violence or terrorism or series of directly related incidents (such as a strike and retaliation)
science and discovery news	announcement of a discovery or breakthrough, technological advances, awards or recognition of a scientific achievement
financial news	specific economic or financial announcements (like a specific merger or bankruptcy announcement); reactions to the event; direct impact on the economy or business world)
new laws	announcement of new legislation or proposals, acceptance or denial of the legislation, reactions.
sports news	a particular sporting event or tournament, sports awards, coverage of a particular athlete's injury, retirement or the like.
political and diplomatic meetings	preparations for the meeting, the meeting itself, decisions, outcomes, reactions
celebrity and human interest news	most often involves the death of a famous person or other significant life events like marriage
miscellaneous news	specific events or activities that do not fall into one of the above categories

Table 5: TDT seminal events and examples

Code	Description	Examples
SpecificEvent	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
GenericEvent	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
IndirectEventReference	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.
PeriodicEvent	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
PersonEntity	Apply this code to identify personal names in topics.	President Bush; Sasha Cohen;
PlaceEntity	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
OrganizationEntity	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;
OtherEntity	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
ExplicitDate	Apply this code to identify explicit dates.	1988; June 15, 1991; October 2007; Monday

Table 6: Codebook