Time-Critical Search

Nina Mishra Ryen W. White Samuel Ieong* Eric Horvitz
Microsoft Research
{ninam,ryenw,saieong,horvitz}@microsoft.com

ABSTRACT

We study time-critical search, where users have urgent information needs in the context of an acute problem. As examples, users may need to know how to stem a severe bleed, help a baby who is choking on a foreign object, or respond to an epileptic seizure. While time-critical situations and actions have been studied in the realm of decision-support systems, little has been done with time-critical search and retrieval, and little direct support is offered by search systems. Critical challenges with time-critical search include accurately inferring when users have urgent needs and providing relevant information that can be understood and acted upon quickly. We leverage surveys and search log data from a large mobile search provider to (a) characterize the use of search engines for time-critical situations, and (b) develop predictive models to accurately predict urgent information needs, given a query and a diverse set of features spanning topical, temporal, behavioral, and geospatial attributes. The methods and findings highlight opportunities for extending search and retrieval to consider the urgency of queries.

1. INTRODUCTION

Web search may be performed in the context of pressing, time-critical challenges where information is needed urgently. For example, acute medical concerns, such as fainting, seizures, bleeding, heart attack, broken bones, numbness, vertigo, and pain may lead to the urgent pursuit of medical information in support of timely action. In time-critical situations, the value of retrieved information diminishes with delays in action informed by the information [27]. Urgency of action has largely been ignored in information retrieval (IR). If search systems could estimate the urgency of searchers' information needs at query time, they could favor content that assists searchers in performing urgent tasks such as making assessments of situations, engaging appropriate assistance, and performing supportive procedures. Surfacing content that is relevant and understandable in time-critical settings could help people with taking timely action in acute situations, where delays can be costly and even life-threatening.

As we will show, people frequently turn to search engines for assistance with urgent medical matters. Search engines might better help people to perform life-saving procedures while they are waiting for emergency medical services (EMS), especially in remote areas. Also, search engines could help people to understand the urgency of situations. EMS has been found to be underutilized because of uncertainty about the relevance of symptoms [23], economic disadvantage [50], and reliance on self-treatment [9]. As an example, in the case of severe chest pain, only 50% of people engage EMS [10]. By some reports, mortality from heart attacks can be decreased by 25% if therapy is initiated within one hour of the onset of acute symptoms [9]. Offering solutions to such challenges could raise the likelihood of valuable actions being rendered in advance of the arrival of EMS [38].

Time criticality has been studied outside of IR, and researchers have examined time–utility tradeoffs in settings such as trauma care [28], emergency medicine [27, 35], communications [26] and aerospace [25]. Research on various temporal aspects of IR has focused on longitudinal analyses of information needs [48, 11], including predicting future events [45], ranking [46], and query auto-completion [49]. Recent work in search evaluation has focused on the inclusion of information gained over time in determining retrieval system performance [31, 51]. However, to our knowledge none of this research has explicitly considered the urgency of situations in the process search and retrieval.

To our knowledge, we provide the first detailed study of time-critical information needs and urgency in search. We focus on the retrieval of time-critical health information needs because Web search is often used to search for health information [20] and a portion of health searches are urgent in nature, performed to address concerns and challenges arising in real time. We employ both a survey and a large-scale analysis of mobile search behavior to understand the nature of such urgent medical needs in retrieval settings, characterize their occurrence, and develop models to predict urgent needs given a query.

Before continuing, we present an example of a time-critical health search situation drawn from the logs of the Microsoft Bing search engine, used in our study (Figure 1). In the example, the searcher is observed issuing queries about diagnosing a potential stroke. A subsequent search for an

*Currently at Google.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
ACM 978-1-4503-2257-7/14/07..$15.00
http://dx.doi.org/10.1145/2600428.2609613.

8:08 heavy limbs fatigue
8:10 heavy limbs fatigue slurred speech
8:17 stroke in women
8:39 slurred speech heavy limbs
8:42 best emergency room in sugarland
8:44 best hospital in sugarland

Figure 1: Sample time-critical search session.

emergency room suggests that the searcher believes that the situation is urgent.

Behavioral evidence such as that in Figure 1 has been used to model search interests and intentions, especially when including result click information [32, 1]. Using such evidence, combined with geographical, topical, and temporal features, we develop classifiers to predict urgent information needs. Estimating the degree of urgency associated with a search request is a necessary first step for developing search systems to support time-critical information seeking. Studies have shown that results returned by search systems for urgent health scenarios are irrelevant, inaccurate, and do not consider the influence of cognitive load on people facing emergency situations [43, 25]. Given an estimate of urgency, search systems could favor particular types of content (e.g., instructional pages containing clear guidelines) or display instant answers comprising graphics or videos to guide people in taking action quickly. Results that are most appropriate for urgent situations might not be ideal for more relaxed settings. For example, the best material for reviewing cardiopulmonary resuscitation (CPR) in preparation for a course on advanced cardiac life support is quite different than the best content to display to frenetic users trying to understand what to do when an elderly relative is suddenly unresponsive and no heartbeat and breathing can be detected.

We shall next report on a survey performed by crowdworkers on recent handling of emergencies and the experience with using search to support decisions during such emergencies. Then we describe the construction and validation of a machine-learned classifier that can predict time criticality, focusing on the definition of sets of predictive features and the curation of the data set, employing an automated labeling procedure. Using nearly half a million sessions over six months of mobile search activity, we build a classifier with 88% positive precision and positive recall.

2. RELATED WORK

Areas of relevant related work include studies in (a) user behavioral modeling, (b) time-aware information access, and (c) models of time-criticality, including efforts in artificial intelligence. We consider each of these areas in turn.

There has been significant research on analysis of search behavior. Search log data has been used to study how people engage in search activity [55], predict users' next online actions, including the timing of next actions and the influence of timing on query content [37, 15], predict users' future interests [17], and to improve the operation of search engines [32, 1]. The influence of the search situation on information needs and activities has been examined in detail in research on incorporation of notions of broader context in search [29]. Focusing on individual users, modeling short- and long-term searcher interests can also be useful for applications such as search personalization [7, 52].

Contextual signals such as location [5] and search task [40] can be used to more completely model searcher's situations at query time. Research on mobile information seeking has characterized usage patterns [34] or developed support for mobile search [41]. However, these methods employ user location in ranking (preferring proximal resources), rather than using this information in a richer model of search intentions as we do in this research. Just-in-time information retrieval agents [47] proactively retrieve information based on user context and alert people to urgent information. We focus on people seeking to satisfy urgent information needs rather than on identifying valuable recommendations.

Studying longitudinal trends in online search behavior within and between users can be useful for understanding the in-world activities of searchers [48, 11], with applications in health care [22, 57]. A recent study explored perceived time pressure on search satisfaction, and search behavior in controlled settings [14]. Time is also emerging as an important aspect of retrieval evaluation, with methods focused on the extent of the information gained over the course of the search session [31, 51].

Time-dependent actions and outcomes have been studied in the artificial intelligence (AI) community, including efforts with representation and reasoning, and applications in automated and human decision making. Studies in the latter domain have focused on the development of methods for the real-time control of the configuration and quantity of information on displays [25], methods for triaging communications [26], and methods for representing and reasoning about time-dependent utilities of action, with applications that include providing decision support for emergency medicine [27] and trauma care [28]. Beyond applications, the work has explored key principles of time-critical decisions in light of computational and cognitive constraints, e.g., [24]. Such careful exploration of the tradeoffs between time and utility could enable search engines to consider degrees of urgency and information value. Human factors researchers have studied how people use information in such time-sensitive work settings as emergency response [35]. Psychologists have also examined the influence of urgency on human behavior [36] and work task performance [8].

Some limited support is available in search engines for particular high-stakes queries (e.g., searches for poison or suicide being linked to hotline telephone numbers [13]). Rather than such sparse special-case linkages, we construct predictive models that estimate the urgency of queries via analysis of query behavior and situational features.

3. MOTIVATION

A review of logs of searches performed in mobile settings obtained from consenting users of the Microsoft Bing search engine reveals a surprisingly large number of queries that could be time critical. However, we cannot confirm urgency in the absence of ground truth. To better understand urgency in search settings, we conducted an initial survey. The goal was to estimate the frequency with which people turned to search engines when faced with time-critical needs, the general circumstances of the situations, and to understand how the degree of urgency influenced the use of search engines.

Table 1: Distribution of recent urgent problems recalled and those who had used a search engine for assistance.

All Respondents	Fall (11%), Breathing Problems (10%), Abdominal Pain (8%), Chest Pain (7%), Allergies (7%), Overdose (6%), Bleeding (6%), Other (6%), Back Pain (5%), Headache (4%), Psychiatric (4%), Eye (4%), Unconscious (3%), Pregnancy (3%), Heart (3%), Diabetes (2%), Convulsions (2%), Choking (2%), Cardiac (2%), Burns (2%), Animal (2%), Stroke (1%), Trauma (1%), Entrapment (1%), Assault (1%)
Search Respondents	Abdominal Pain (14%), Breathing Problems (10%), Other (10%), Allergies (9%), Chest Pain (7%), Fall (5%), Eye (5%), Overdose (5%), Heart Problems (4%), Headache (4%), Burns (4%), Bleeding (4%), Back Pain (4%), Psychiatric (3%), Pregnancy (3%), Diabetes (3%), Cardiac (2%), Animal (2%), Unconscious (1%), Convulsions (1%), Choking (1%), Trauma (1%), Heat Cold (1%), Electricity (1%)

For the survey, we employed Amazon's Mechanical Turk (AMT) to recruit a set of crowdworkers drawn from the U.S. population. Two surveys were conducted. The first was aimed at understanding general characteristics of how users behave in an emergency. The second survey focused specifically on the use of search in emergencies. The sample comprised of AMT Masters, workers who have completed at least 1000 hits with an approval rating >95%. The surveys had 133 (120) respondents. The distribution of gender was 50% F, 50% M in the first survey and 46% F, 54% M in the second. In both surveys, we only considered participants who had recalled being in an emergency situation within the past year, so as to ensure that details of the situation were easier to remember. Several biases have been noted in the general AMT population demographics (e.g., younger, female, seeking auxiliary income [44]), and these biases should be considered when interpreting the results.

The results indicate that 12% of users turn to a Web search engine in an emergency. Among those with a smartphone, 16% turned to search for additional assistance. Time-critical sessions constitute a small fraction of all search traffic. However, the high stakes of action in time-critical situations motivate efforts to detect and respond to urgency.

One goal of the survey was to understand the types of urgent problems experienced among all users vs. those who used a search engine. To gain insights about the distribution of urgent problems , we asked participants to categorize their time-critical needs using a classification available from the Medical Priority Dispatch System [12], a U.S. national standard for 911 response coding. The distribution of problems reported by users is shown in Table 1. On the first line, we show the distribution among all respondents (first survey) and on the second line only those who had used a search engine to pursue information on the urgent matter at hand. Some differences are noticeable. While falls are the most common category among all respondents, they are less likely among those who used a search engine. A possible explanation is that people better understand actions appropriate for helping with falls better than courses of action appropriate for many other scenarios. The seriousness of a situation, such as the onset of severe abdominal pain, and ideal actions for addressing a concern may be more confusing. For example, "I try to see what was caus my abdominal pain but couldn't really come up with anything" and "Since my pain was on my lower left side of my body, I did some Google searches on symptoms of appendix problems" (we note that pain from appendicitis is typically experienced on the right or symmetrically across the abdomen).

The respondents' descriptions of the use of search engines shows a strong reliance on search. In some cases, online search precedes a visit to the emergency department (ED) to access information on what the problem might be and/or what action should be taken. Here is an example typical of the use of search as first responder: "My father had a lot of pain in his left arm, and I looked up what it could be. It turned out he was having a mild heart attack based on my results we figured we should call 911." One person said "I googled what to do after someone struggled swallowing food" when his friend was choking on dinner. He performed the Heimlich Maneuver, and brought the friend to the hospital thereafter. In other situations, search engines are used to guide searchers through the steps of the treatment process, e.g., "I witnessed someone having a heart attack. I had to start cpr and had forgotten how. I quickly searched the cpr technique." In other cases, the searches follow a call to 911. For example, a user whose child was having convulsions searched for [first time seizure in toddler] and [seizure first aid] after calling 911. Another user "I searched head injuries to see if there was anything else I could do while waiting for the ambulance. I searched: head injury immediate treatment."

Among those who used search engines, 83% were satisfied and 17% were not satisfied with the information they had retrieved. Among the satisfied, some explained that they could not blame the search engine since their symptoms could mean the presence of many different conditions, e.g., there are many causes of abdominal pain. Those who were not satisfied with search provided a variety of complaints. The quality of displayed results were called into question by some, e.g., a participant whose girlfriend was hit by a car while walking mentioned, "I just couldn't find anything that I thought was reliable, well written information. I was also probably in a slight state of shock, so simplified articles would have been easier to digest." Others complained about advertisements,"The information that came back in bulk was mostly junk sites looking to earn revenue from ads. No real medical information. The couple of sites that did have okay info, such as webmd, were so wishy washy in answers due to litigation possibilities that it rendered their info worthless too. I finally found some good peer reviewed info. Good thing I know the medical jargon." Others complained about the quantity of search results "there was just to much stuff that came up, in a situation when you are starting to panic you get overwelmed really quick, especially when you think you might be having a heart attack." Other participants described being moved to anxiety by results that turned out to be erroneous, e.g., "Less dire results. It kept telling me

Table 2: Top categories of a random sample of positive and negative training data.

MeSH Positive Category (%)	Sample Queries from Positive Set
Bacterial Infections (18%)	[symptoms of appendicitis], [what do u do when u have strep]
Virus Diseases (12%)	[can you get shingles more than once], [pregnant flu symptoms]
Pathological Conditions (12%)	[sudden upper abdominal pain], [kidney stone signs]
Respiratory Tract Diseases (8%)	[pneumonia in the elderly], [pneumonia symptoms 1 year old]
Wounds and Injuries (6%)	[broken toe what to do], [sudden bruising in foot]
Pregnancy Complications (6%)	[having an iud and miscarrying symptims],[pregnancy calculator]
MeSH Negative Category(%)	**Sample Queries from Negative Sessions**
No Health Query (15%)	[what makes blood boil], [heart of glass], [protein shakes]
Reproductive Phenomena (10%)	[pregnancy calculator week by week], [can you ovulate twice in one month]
Organic Chemicals (8%)	[cexedrine], [azithromycin 250 mg], [divalproex er], [tylenol pm recall]
Pathological Conditions (7%)	[weight loss patch], [ingrown toenail]
Neoplasm (7%)	[esophageal cancer], [skin cancer], [signs of bone cancer in leg]
Heterocyclic Compounds (7%)	[what is melatonin usex for], [how long does hydrocodone stay in your system]

I was probably having a heart attack just because I was having what felt like chest pain but I wasn't."

Summary of Survey Findings The survey showed that 12% of users turn to search engines for time-critical needs and 16% among smartphone users among our respondents have done so. We see from the responses that people are using search for symptom searches in pursuit of diagnoses and suggestions for what to do–and that there is somewhat of a mismatch between general urgency and urgent search needs. The key findings are: (1) people rely on search engines for urgent needs, (2) search is often used as a first response in emergency situations, and (3) people are dissatisfied with search results in terms of both quality and quantity of returned results, with advertisements appearing, and with the overall complexity of content. Many of these challenges could be met by search engines if they had the capability to understand the degree of urgency of users' information needs, and could tailor the search experience accordingly.

The challenges and opportunities raised in the survey frame a larger research agenda on time-critical search, including efforts to detect the time-criticality surrounding a query, retrieving appropriate content, and the identification, construction or manual curation of content designed for assisting with action in urgent settings—when people are likely under significant duress and cognitive load. We focus in this paper on one step of the process: on identifying when a user likely has a time-critical need. This triggering component is a crucial, early step in the pipeline of providing useful, timely help to users. We leave the task of finding or constructing useful time-critical content as a future direction.

4. PRELIMINARIES

Search engine logs have been used to understand search behavior [33, 55], predict future activity and interests [37, 15], improve search engines [32, 2], and learn about the world [48, 11]. A *search session* is a contiguous sequence of queries from a user with a short time gap between searches [15].

As a means of tagging for time criticality, we assume in our study that a user has a *time-critical health information need* if they seek an urgent care facility during the search session. A user is noted as seeking assistance from an urgent care facility if they issue a query, call by phone, or seek driving directions to an emergency department.

Per the example in Figure 1, while we cannot be certain, this sequence appears to represent someone trying to diagnose a possible stroke. The search for an emergency room suggests that they believe the situation to be urgent. We realize that numerous urgent health situations do not lead to such searches or calls to an emergency department. For example, information retrieved about the Heimlich Maneuver may be used to assist someone with choking without resulting in a follow-on pursuit of professional medical assistance. Conversely, people may seek out emergency care in cases where it is not needed. For example, people with unfounded anxiety about their health [56] may search for an emergency room when such medical attention is not necessary. Nevertheless, we propose that searches for emergency medical assistance are dominated by *user-perceived* time-criticality. We explore this further in Section 5.

We seek to build and evaluate predictive models that can identify a user in a time-critical situation. More specifically, given a user's search history and current search session, we seek to predict if they are in a time-critical situation. To do so, we employ supervised machine learning to develop a classifier, leveraging sets of features and labeled data derived from in situ search behavior.

5. TRAINING DATA

We now describe how we automatically generate training data, i.e., positive and negative time-critical sessions. To better understand the nature of this data and to help verify the utility of our method for automatically generating time-critical labels, we also characterize a small sample into an existing medical taxonomy. We describe each of these steps in the remainder of this section.

5.1 Automatic Label Generation

We opt for the automated labeling of user sessions given related challenges found with using human judges to assess relevance [2], and the difficulty that we believe that human annotators may experience in reliably labeling true urgency solely from search queries. However, we cannot assume that such phrases as "emergency room" indicate a time-critical need, as such phrases can be used in other contexts, e.g., the television show ER. Broader contextual information, such as follow on searches can help with the discrimination task. These challenges motivate the following method:

1. Identify sessions containing health queries.

2. Remove sessions containing adult content.
3. Among the remaining sessions, if a session contains a search or call or directions for emergency room, and the session ends or the topic of the session after this search does not shift away from health, then the session is labeled as positive for time criticality. Otherwise, the session is labeled negative.

Step (1) tends to eliminate non-health information needs. Step (2) is a filter to improve data quality. Step (3) eliminates sessions that switch away from the health, demonstrating that the searcher is not consumed by the situation and that it may not truly be urgent.

5.2 Characterizing Training Data

To verify that the remaining sessions are labeled correctly and to glean insights about the operation of the automated labeling procedure, we randomly sampled 50 examples of positive and negative cases from the tagged set. We manually categorized these sessions into the Medical Subject Headings (MeSH) hierarchy produced by the National Library of Medicine [39]. The top categories of the positive and negative sessions are shown in Table 2.

The top positive sessions are generally of a more urgent nature than the negative sessions, including queries that seek such information on broken bones, pneumonia in children, and diagnosis of appendicitis. However, individual queries are not sufficient for distinguishing time-critical from non-urgent settings. For example, [pregnancy calculator] appears in both positive and negative data, but the surrounding context is different. In the positive set, the query was posed in an Ale House, while in the negative data, the query was posed in a residential area. We revisit the broader query context, considering such factors as location, in Section 6.

Among the positive cases sampled, we found that one session was clearly non-urgent (user seeking an emergency center for Vicodin to evade a drug test for a construction job). Other sessions labeled as positive include those where we are uncertain about the true urgency of a situation (e.g., emergency room needed for strep throat). However, we note that our goal is to predict when users believe they have an urgent need and seek urgent care. We do not intend for search to take on the role of an expert diagnostician.

The top five categories of the negative examples are also shown in Table 2. We note that a large subset of sessions do not include health-related queries, i.e., the health query categorizer mistakenly fires. Such queries may contain health-related terms, such as [what makes blood boil] and [protein shakes]. Ideally, the health query categorizer would not make such mistakes. Nevertheless, these sessions truly are negative cases, so the time-critical classifier should learn to label these sessions negative.

Among the categories that fall into the MeSH hierarchy, the largest negative category is *Reproductive Phenomena*. This category contains queries such as [pregnancy tests]. Among the remaining categories, medication-related searches are common, e.g., [tylenol] and [melatonin]. Cancer-related sessions are also common.

While the category *Pathological Conditions* is common to both the positive and negative training data, we found that the queries in the positive examples are of a more acute nature. The positive set contains [kidney stones], [chest pain], and [throwing up foam] while the negative data contains [weight loss patch] and [ingrown toenail].

6. FEATURES

To identify a query as urgent, we defined several observational features. These features include characteristics of the user, historical behavior of all previous users who issued the query, words in queries observed thus far in the session, situated features related to topical and behavioral aspects in the session, as well as temporal and geospatial characteristics. A full description of these features can be found in Table 3. We now explain the rationale behind the selection of features.

6.1 User Characteristics

Statistics published by the Centers for Disease Control (CDC), based on ED visits between 1997-2011, show that 20% of adults visited an ED in the 12-month period, and 7% visited the ED twice or more in that period [19]. By studying a user's longitudinal search behavior, we can better estimate the likelihood that their current need is urgent. For example, if a user already displays evidence of seeking urgent medical attention (e.g., via the presence of searches for professional health care) in the time period preceding the current health search, it may be less likely that they have an urgent health need related to the current query.

Other features of searchers' longitudinal search behavior may be useful for predicting an urgent need. For example, if a searcher typically issues long queries but the average query length in the current session is extremely short, the reduction in length may indicate something about the atypical urgency of information needs at the current time.

6.2 Historical Query Statistics

Query statistics about the query and browsing behavior of populations of searchers provide additional information. For example, the timing and nature of query refinements and clickthrough information on search results can provide insights about the adequacy of results. To this end, we include several historical query statistics for current health query in our analysis, including the frequency with which the query was received by the search engine and the corresponding clickthrough rate on search results.

6.3 Query Words

The terms appearing in the queries observed in the session provide some insight into the nature of searchers' information needs. The presence of words such as [help] provide direct evidence of urgent information needs. However, more subtle clues about time criticality may be present in a user's query stream, e.g., the mention of symptoms such as [chest pain] or [choking], coupled with the presence of other terms such as [tightness] or [trouble breathing]. Rather than specifying the words in advance of the prediction, by including the words directly as a feature, our models have the opportunity to learn which terms are important. Queries triggered by the automatic labeling scheme, e.g.,[emergency room] were not included in the bag of words as the existence of such terms would determine the label of the session. We use the query words as a baseline in the prediction experiments presented later in the paper.

6.4 Situated Features

Features describing the current situation of the user could be most telling about whether the information being sought is urgent in nature. To capture aspects of the users' context,

Table 3: Features used in predicting urgent information needs

Feature Type	Description of Features [Number of Features]
Bag of Words	Words in the query (unigrams) [Undersampling: 2,132, Realistic: 56,429]
Historic Query Statistics	Number of times the query was issued, Average number of ad clicks per query instance, Average number of result clicks per instance, Average dwell time of result clicks, Average time to first result click, Average time to last result click, Average number of successful clicks (dwell $\geq$ 30s), Average number of switches to other engine [43 – includes additional if null features]
User-Related (preceding current session)	Number of queries from current user, Number of sessions from current user, Number of days with a query from current user, Number of days with a query from current user, Number of time-critical facility queries from current user, Number of calls or directions from current user, Average query length for current user [7 features]
Situated	[43 features]:
Behavioral	Number of queries, Average term overlap between consecutive queries, Average length of queries, Average time between queries, Number of queries without result clicks, Number of queries with result clicks, Number of result clicks, Number of unique Web domains for result clicks, Duration
Geospatial	Distance traveled, Average speed of travel, Altitude, Distance to nearest {urgent care facility, hospital, park, mall, school, pitch (soccer field), recreation ground, body of water, retail store, kindergarten, sports stadium, running track}
Temporal	Current query time in three hour bucket (e.g., 12PM-3PM), daylight (Between sunrise and sunset in the location of query at day of year), working hours (hospital open), weekend, Saturday, Sunday
Topical	Number of unique topics (based on top result for query), Number of topic changes between consecutive queries, Number of health queries, Number of new (not in user history) topics, Health query is new topic for user, Change in topic at health query (versus previous topic)

we developed a number of classes of features to represent behavioral, topical, temporal, and geospatial aspects of the user's current situation. We refer to these observations as *situated* features.

Behavioral Aspects of user search behavior may reveal additional distinctions about information needs. Search behaviors, such as the average number of clicks associated with a particular query, have been used in previous research to understand the nature of the search intent, e.g., to distinguish between informational and navigational search queries [53]. We hypothesize that search behaviors may also be useful in distinguishing urgent from non-urgent needs. For example, urgent needs may manifest as a rapid sequence of overlapping short queries, and less urgent queries may be characterized by longer inter-query times and more clickthrough on content as searchers have the time to examine multiple results in detail. To this end, we consider such attributes as the number of queries issued in the session so far, the average time between those queries, and the number of search results that were visited by searchers.

Topical We believed that the level of fluctuation in a user's topical interests and the sudden emergence of new topics could by triggered by an urgent need. Topic-based features capture the dynamics of interests in the session so far, as well as the relationship between the topics that the user is interested in and those observed in their long-term search behavior. We hypothesize that, because urgent events are rare, the emergence of a new topic (e.g., the initiation of a health search) would reveal something about the rarity of those events and increase the likelihood that an unforeseen health event is occurring. We also consider the dynamism of the searcher's information needs within the session, e.g., to what extent are their topical interests stable (including the past history of health queries), and whether there is a change in topical interests at the point in time where health queries assume the focus of attention.

We assigned topic labels to queries by automatically classifying the content of the top-ranked search results returned for the search engine for that query. If no results were returned by the engine, no label was assigned. Topic labels were taken from the 219 topics from the top two levels of the Open Directory Project (ODP, http://dmoz.org), and included topics such as "Health/Medicine" and "Recreation/Sports". The topics were assigned to pages based on their content using a text-based classifier described and evaluated in [6]. The coverage of the classifier across all result URLs in our test set was 96.8%; the remaining 3.2% of URLs were unreachable at feature generation time.

Temporal Circadian rhythms play a role in people's health and well-being, and people are more prone to injury at certain times of the day. For example, the most dangerous times for heart attack and for other cardiovascular emergencies, including sudden cardiac death, rupture or aneurysm of the aorta, pulmonary embolism and stroke are the morning and during the last phase of sleep. A recent study showed that there is a threefold increase in the risk of myocardial infarction in the three hours after waking, related to oscillations in the cardiovascular system [42]. Time of day may also be indicative of the degree of surprise in a rising health concern. A health query arriving at 2:00am may be more remarkable (and indicative of an unforeseen health event) than a health query arriving in the middle of the day.

Time also influences the nature of the activities that people are likely to be performing. During business hours (9AM-5PM) on weekdays, people may be engaged in sedentary and safe activities in office environments. In the evenings and on weekends people may more typically pursue other interests, bringing them into situations with higher risk of injury and

of placing additional strain on their bodies—and creating opportunity for unforeseen accidents. Indeed, examining the positive examples in our data as a function of time-of-day and day-of-week, we observe a greater likelihood of urgent health searching occurring outside of working hours and on weekends (Table 4). Note that this differs from when emergency rooms are more likely to receive visits [18], suggesting that urgent search engine temporal patterns may differ from ER visit patterns.

In addition to the absolute time at which a query is issued, we also consider whether the health search of interest is conducted during daylight hours. Hours of darkness vary depending on geographic location as well as time of year, and the presence of darkness has been shown to affect injury likelihoods; accident statistics have shown that fatal vehicular accidents are three times more likely at night than during the day [54].

Geospatial Location may provide strong evidence about the nature of the searcher's current situation [3]. We compute proximity to locations where we suspect people might be engaged in physical activity (parks, pitches, running tracks, and recreation grounds) and locations where large numbers of people are likely to gather (e.g., at stadiums or shopping malls) or EMS is hard to reach. Since children may be particularly susceptible to injury, we also use the distance from the location of the query to the nearest school. Since speed of motion might reveal something about urgency, we include both the current velocity and distance traveled since last search query as features.

In addition to considering the activities that people are likely to perform, we also consider the likelihood that they could readily receive medical attention at their current location. We compute the distance from query location to the nearest urgent care center, including EDs of hospitals. If a searcher is near a medical facility, one might argue that the query itself is less likely to represent a genuine cry for help as medical attention is immediately available. We also consider other details about the surrounding environment. We consider whether the query was issued while the user was over water (user is likely on a boat) and the altitude of the user at query time (e.g., if high, user may be hiking or climbing). Both of these features provide evidence of difficulty in reaching medical care soon and could influence searchers turning to their mobile devices to seek assistance so as to guide decisions about a course of action in advance of receiving professional medical assistance.

7. INFERRING TIME CRITICALITY

We develop a classifier that can distinguish between the positives and negatives using the set of features described above. As the data is heavily skewed to negatives, a classification policy that always predicts negative would have a high classification accuracy. For our initial study, we sample a smaller set from the negatives, so that we have an equal number of positives and negatives [30]. The benefit of such downsampling is that the performance of a random guessing baseline is 50%. Such an event split enables us to tease out the effects of including various features.

While the study of classification constructed with balanced data can build insights, that target is use in a realistic setting. Our second set of experiments probe the value of the methods in a realistic scenario: we train on one month of data and test on the following month with actual statistics as seen in logs. Particular emphasis is given to positive precision and positive recall. We also demonstrate that accuracy improves as the amount of session data used for feature generation increases.

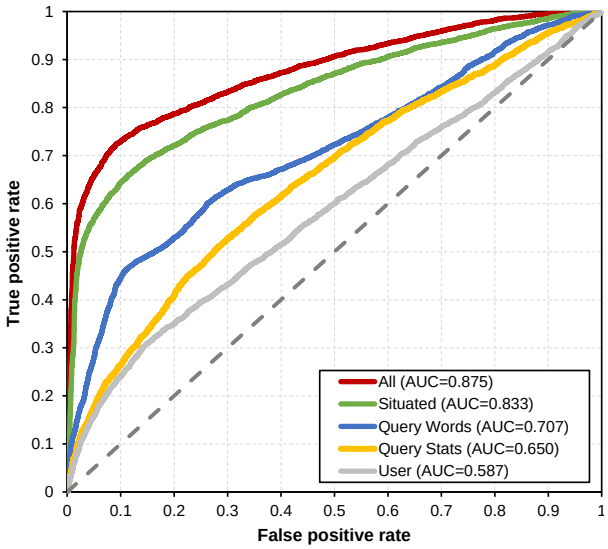

Figure 2: ROC curves comparing the performance of classifiers trained on user features, historic query statistics, words in the query, situated signals and all queries.

7.1 Experiment Setup

We collected six months of search activity from Bing mobile search spanning 9/2012 through 2/2013. The application is typically installed on tablets and mobile devices. The mobile logs contain search queries, clicks, timestamp, and real-time latitude and longitude users who consented to share data when setting up the application. As expected, the data is severely skewed towards the negative examples; we found 536 positive examples and 945,989 negative examples (99.4% negatives and 0.6% positives).

In our experiments, we train a boosted tree classifier [21]. We perform a 10-fold cross validation, repeating each 10 times. Results are reported in terms of precision and recall. Positive precision is the fraction of sessions predicted time-critical that actually are time critical. Positive recall is the fraction of actual time-critical sessions that are predicted time-critical (and similarly for negative precision and recall).

7.2 Downsampling

Our first experiment uses balanced data, constructed via downsampling. We assess the predictive power of different categories of features by separately training classifiers with distinct sets of features per the categories of features in Section 6, i.e., words in the query, user features, historic query statistics, situated only features, and all features. We assess each of the categories separately to understand their intrinsic benefit. An alternative is to perform feature ablations, but that depends highly on the order in which features are removed and is affected by interactions between feature categories.

The classical approach for text categorization [16] involves representing the query as a bag of words. Phrases are not

Table 4: Distribution of day of week and time of day when positive and negative sessions take place. Positive sessions are more likely on the weekends and night time/early morning.

Day of Week	Positive (%)	Negative (%)	Difference	**Time of Day**	Positive (%)	Negative (%)	Difference
Monday	14%	15%	-1	12am-3am	16%	13%	+3
Tuesday	12%	15%	-3	3am-6am	17%	15%	+2
Wednesday	11%	15%	-4	6am-9am	16%	16%	0
Thursday	13%	15%	-2	9am-12pm	18%	18%	0
Friday	13%	13%	0	12pm-3pm	10%	20%	-10
Saturday	19%	13%	+6	3pm-6pm	8%	9%	-1
Sunday	18%	14%	+4	6pm-9pm	7%	3%	+4
				9pm-12am	8%	6%	+2

Table 5: Evaluation of the predictive power of different feature subsets obtained via downsampling the negative examples. Numbers are averages (standard deviation) over 10 runs.

Features	Accuracy	Positive Precision	Positive Recall	Negative Precision	Negative Recall	Area Under Curve
Bag of Words	0.665 (0.044)	0.706 (0.057)	0.574 (0.068)	0.639 (0.040)	0.757 (0.059)	0.712 (0.044)
Query Statistics	0.610 (0.047)	0.606 (0.055)	0.590 (0.069)	0.618 (0.044)	0.631 (0.081)	0.650 (0.054)
User-related	0.556 (0.044)	0.562 (0.045)	0.551 (0.066)	0.551 (0.045)	0.562 (0.066)	0.587 (0.047)
Situated	0.771 (0.042)	0.806 (0.057)	0.697 (0.057)	0.749 (0.038)	0.841 (0.056)	0.834 (0.040)
All	0.815 (0.042)	0.846 (0.059)	0.745 (0.068)	0.797 (0.044)	0.877 (0.054)	0.876 (0.040)

known to add predictive power over words [4], so we focus on words only. Since our underlying data comprises search sessions, our bag of words contains all queries up to and including the first health query of a session. As our goal is early detection of time criticality, we seek to detect time-criticality after the arrival of the first health query. The overall accuracy of the bag-of-words classifier is 66% and a detailed breakdown by precision and recall is displayed in Table 5.

Next, we experimented with historical query statistics. These features are based on the activity trails of users who had issued the query. A categorizer built exclusively on these features is 61% accurate. The features with the most predictive power are average amount of time spent on clicked search results, average time to first click, and features related to switching search engines. All of these features capture a user in a more panicked state, i.e., dwelling less, clicking faster, and quickly shifting to another search engine in hope of finding better content.

We found that user-related features have little predictive power in themselves. The overall accuracy of classifiers based solely on user-related features is 55%, which is only marginally better than random guessing 50%. Such a result is not unexpected as relying solely on a user's behavior preceding the current session is not likely to contain enough information to predict time-criticality on a future emergency.

Harnessing situation-related features (topics, location, altitude, proximity to sites and resources, and time of day) enables classification of time criticality with a classification accuracy of 77%, with significant boosts in positive and negative precision and recall. Digging deeper into these *situational* features, the most predictive are session-related and include the number of new and unique topics by the user. Other situational features are related to what we can infer from a user's location, specifically distance to closest stadium, hospital, park, mall, retail store, and school. These features suggest that the user is away from home, e.g., attending an event in a stadium or park, shopping at a mall or store, or in a playground at school. The next highest-ranked feature by predictive power is altitude. We could identify time-critical sessions in mountainous regions where a mobile device may be the only source of information. Note that situational features perform better than bag of words despite the fact that there are many more word features (2132 words vs. 43 situated).

Finally, we trained a classifier on all of the above features and achieved an accuracy of 82%. It is striking that the situational features provide a great deal of the predictive power of the classifier. The exact words in the query do not matter as much as the circumstances of the user. Figure 2 shows a receiver-operator characteristic (ROC) curve comparing the performance of classifiers based on (1) user, (2) query statistics, (3) bag of words (4) situational, and (5) all features. All paired differences in differences in accuracy are strongly statistically significant using a t-tests (p-value < 0.001).

7.3 Realistic Setting

In a more realistic scenario, we can train and test on a larger set of negative data. In the next experiment, we train on sessions collected in January 2013 and evaluate on sessions collected in February 2013. The breakdown of data in January 2013 is 139 positives, 230K negatives and February 2013 is similar 136 positives, 238K negatives. The positives are needles in a haystack of negatives. Although they are rare, they may represent an extremely critical information need. Using all the features, we now compare our ability to predict time criticality after the first and after the last health query in a session. The goal is to show that even if we miss a time-critical session after the first health query, we can still flag it as the session progresses. Note that we must be careful not to use words later in the session that contain time-critical facilities such as "emergency room,âĂİ since that is how we labeled sessions positive. Thus, queries containing time-critical facilities are not included in the bag of words. Session features can get richer as the session progresses, so there is potential for improving prediction performance by employing features beyond the first health query, including

Table 6: **Prediction results for realistic setting.**

Session-depth	Accuracy	Positive Precision	Positive Recall	Negative Precision	Negative Recall	Area Under Curve
First health query	0.999	0.875	0.883	0.999	0.999	0.999
Last health query	0.999	0.888	0.928	≈ 1	0.999	0.999

such behavioral information as whether a user shifted to a non-urgent topic.

The results for the realistic setting are shown in Table 6. Total accuracy, as well as negative precision and recall, are not important in this experiment, as a classifier that labels every example negative will have good performance for these measures. Instead, we attend to the positive precision and positive recall. Positive precision (88%) is higher than seen in the experiment with downsampled data (82%). However, there is not much difference in positive precision between the first and last health query on all data. Conversely, positive recall improves with using information up to the last health-related query (from 88% to 93%); more session data means more opportunity to detect time criticality.

8. DISCUSSION AND IMPLICATIONS

We provided evidence that people turn to search engines for urgent needs. Via surveys and large-scale search log analysis, we quantified and characterized urgent behavior. We identified a set of novel predictive features, especially those that are situational in nature, including location, altitude, proximity to specific sites and resources, and time of day. Using these features, we developed a machine learning classifier that can provide accurate predictions about the level of urgency associated with search sessions. These predictions can be used by search systems to determine whether specialized ranking or interface support should be offered.

Our work has limitations that are important to consider. First, as mentioned earlier, the AMT population is known to be biased in a variety of ways [44]. We asked crowdworkers for their gender and state of residence to determine if our samples were overly skewed. However, we did not ask for income and other potentially relevant demographics. In addition, search engine logs inherently reflect a biased population of the US. For example, we may be missing signals from people who do not have access to computers or search engines. Finally, the training data that we generate is also imperfect: a user may search for an emergency room and not really be in an urgent state and users facing a time-critical situation may not search for or call a nearby emergency room. To address this, more work is needed on methods to attain ground truth about urgency from logged searchers, perhaps by constructing pools of consenting searchers who are willing to share additional details about their search situations in addition to their logged search activity.

Our findings have several implications for search engines and IR more broadly. Search engines should consider estimated urgency as an important aspect of the retrieval. Recognizing urgency of a session or an individual search query is only a first step in a larger pipeline of time-critical analysis, content generation, and information display. Multiple designs for assistance are feasible and need to be studied. These designs extend beyond IR. For example, if search engines can determine that a user has an urgent need, actions such as notifying emergency services might be taken.

9. FUTURE WORK AND DIRECTIONS

We believe that time-critical search is important and underexplored. By coupling machine learning along with defining a set of rich situational features, we showed that time-critical queries can be detected automatically. The findings of this study are promising. Nevertheless, much work remains on developing and evaluating retrieval technologies to support searchers faced with urgent, emergency situations. One of the key challenges is providing users with access to content that is easy to comprehend and is actionable in real-time, considering distractions and cognitive load in urgent situations. Even relevant pages may contain complex and lengthy textual descriptions and obfuscating advertisements and graphics. The automatic identification of accurate, succinct, and visually compelling videos or easy-to-digest graphics on procedures could be of tremendous help to people with urgent needs. In addition, the mobile phone provides additional sensors (location, microphone, camera, accelerometer) that could assist a user beyond web pages. We believe that there is great promise in combining methods for identifying urgent information needs with methods for identifying or generating content – as well as using the phone as a sensor – to develop an end-to-end approach at assisting searchers in time-critical settings.

Acknowledgments

We thank Isabelle Stanton for early help with the survey.

10. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proc. SIGIR*, pages 19–26, 2006.

[2] R. Agrawal, A. Halverson, K. Kenthapadi, N. Mishra, and P. Tsaparas. Generating labels from clicks. In *WSDM*, pages 172–181, 2009.

[3] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *WWW*, pages 357–366, 2008.

[4] R. Bekkerman and J. Allan. Using bigrams in text categorization. Technical report, Department of Computer Science, University of Massachusetts, Amherst, 2004.

[5] P. Bennett, F. Radlinski, R. White, and E. Yilmaz. Inferring and using location metadata to personalize web search. In *Proc. SIGIR*, pages 135–144, 2011.

[6] P. Bennett, K. Svore, and S. Dumais. Classification enhanced ranking. In *WWW*, pages 111–120, 2010.

[7] P. Bennett, R. White, W. Chu, S. Dumais, P. Bailey, F. Borisyuk, and X. Cui. Modeling the impact of short- and long-term behavior on search personalization. In *Proc. SIGIR*, pages 185–194, 2012.

[8] D. Bingham and B.J. Hailey. The time-urgency component of the type a behavior pattern. *J. Applied Social Psychology*, 19:425–432, 1989.

[9] A.L. Brown, N.C. Mann, M. Daya, et al. Demographic, belief, and situational factors influencing the decision to utilize emergency medical services among chest pain patients. *Circulation*, 102(2):173–178, 2000.

[10] J.G. Canto, R.J. Zalenski, J.P. Ornato, et al. Use of emergency medical services in acute myocardial infarction. *Circulation*, 106(24):3018–3023, 2002.

[11] H. Choi and H. Varian. Predicting the present with google trends. *Economic Record*, 88(s1):2–9, 2012.

[12] J. Clawson. Medical priority dispatch system. In *Wikipedia*, 2014.

[13] N. Cohen. 'Suicide' query prompts Google to offer hotline. In *NYT*, April 2010.

[14] A. Crescenzi, R. Capra, and J. Arguello. Time pressure, user satisfaction and task difficulty. In *ASIST*, 2013.

[15] D. Downey, S. Dumais, and E. Horvitz. Models of searching and browsing: Languages, studies, and application. In *Proc. IJCAI*, pages 2740–2747, 2007.

[16] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proc. CIKM*, pages 148–155, 1998.

[17] G. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *Proc. SIGIR*, pages 331–338, 2008.

[18] National Center for Health Statistics. National hospital ambulatory medical care survey: 2006 emergency department summary. In *Center for Disease Control and Prevention*, 2008.

[19] National Center for Health Statistics. Special feature on emergency care. In *CDC and Prevention*, 2012.

[20] S. Fox and M. Duggan. Health online 2013. *Pew Internet and American Life Project*, 2013.

[21] J. Friedman. Greedy function approximation: a gradient boosting machine. *A. of Statistics*, pages 1189–1232, 2001.

[22] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.

[23] R. Horne, D. James, K. Petrie, J. Weinman, and R. Vincent. Patients' interpretation of symptoms as a cause of delay in reaching hospital during acute myocardial infarction. *Heart*, 83(4):388–393, 2000.

[24] E. Horvitz. Principles and applications of continual computation. *Artificial Intelligence J.*, 126:159–196, 2001.

[25] E. Horvitz and M. Barry. Display of information for time-critical decision making. In *UAI*, pages 296–305, 1995.

[26] E. Horvitz, A. Jacobs, and D. Hovel. Attention-sensitive alerting. In *Proc. UAI*, pages 305–313, 1999.

[27] E Horvitz and G Rutledge. Time-dependent utility and action under uncertainty. In *UAI*, pages 151–158, 1991.

[28] E. Horvitz and A. Seiver. Time-critical action: Representations and application. In *Proc. UAI*, pages 250–257, 1997.

[29] P. Ingwersen and K. Järvelin. *The turn: Integration of information seeking and retrieval in context*, volume 18. Springer, 2005.

[30] N. Japkowicz. The class imbalance problem: Significance and strategies. In *International Conference on Artificial Intelligence*, pages 111–117, 2000.

[31] K. Jarvelin, S.L. Price, L.M.L. Delcambre, , and M.L. Nielsen. Discounted cumulated gain based evaluation of multiple-query ir systems. In *Proc. ECIR*, pages 4–15, 2008.

[32] T. Joachims. Optimizing search engines using clickthrough data. In *Proc. KDD*, pages 133–142, 2002.

[33] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *TOIS*, 25(2):7, 2007.

[34] M. Kamvar and S. Baluja. A large scale study of wireless search behavior: Google mobile search. In *Proc. SIGCHI*, pages 701–709, 2006.

[35] J. Landgren. Making action visible in time-critical work. In *Proc. SIGCHI*, pages 201–210, 2006.

[36] F. Landy, H. Rastegary, J. Thayer, and C. Colvin. Time urgency: The construct and its measurement. *Journal of Applied Psychology*, 76(5):644–657, 1991.

[37] T. Lau and E. Horvitz. Patterns of search: Analyzing and modeling web query refinement. In *Proc. UM*, pages 119–128, 1999.

[38] E. Lerner, T. Rea, B. Bobrow, et al. Emergency medical service dispatch cardiopulmonary resuscitation prearrival instructions to improve survival from out-of-hospital cardiac arrest. *Circulation*, 125(4):648–655, 2012.

[39] C. Lipscomb. Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[40] J. Liu and N. Belkin. Personalizing information retrieval for people with different levels of topic knowledge. In *JCDL*, pages 383–384, 2010.

[41] D. Mountain and A. Macfarlane. Geographic information retrieval in a mobile environment: evaluating the needs of mobile individuals. *J. Info. Science*, 33(5):515–530, 2007.

[42] J. Muller, P. Stone, Z. Turi, J. Rutherford, et al. Circadian variation in the frequency of onset of acute myocardial infarction. *NEJM*, 313(21):1315–1322, 1985.

[43] K. Murugiah, A. Vallakati, K. Rajput, A. Sood, and N. Challa. Youtube as a source of information on cardiopulmonary resuscitation. *Resuscitation*, 82(3):332–334, 2011.

[44] G. Paolacci, J. Chandler, and P. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419, 2010.

[45] K. Radinsky and E. Horvitz. Mining the web to predict future events. In *WSDM*, pages 255–264, 2013.

[46] K. Radinsky, K.. Svore, S.. Dumais, M. Shokouhi, J. Teevan, A. Bocharov, and E. Horvitz. Behavioral dynamics on the web: Learning, modeling, and prediction. *ACM TOIS*, 31(3), 2013.

[47] B.J. Rhodes and P. Maes. Just-in-time information retrieval agents. *IBM Systems journal*, 39(3.4):685–704, 2000.

[48] M. Richardson. Learning about the world through long-term query logs. *TWEB*, 2(4), 2008.

[49] M. Shokouhi and K. Radinsky. Time-sensitive query auto-completion. In *Proc. SIGIR*, pages 601–610, 2012.

[50] D.B. Siepmann, N.C. Mann, J.R. Hedges, and M.R. Daya. Association between prepayment systems and emergency medical services use among patients with acute chest discomfort syndrome. *Annals of emergency medicine*, 35(6):573–578, 2000.

[51] M. Smucker and C. Clarke. Time-based calibration of effectiveness measures. In *Proc. SIGIR*, pages 95–104, 2012.

[52] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. In *Proc. KDD*, pages 718–723, 2006.

[53] J. Teevan, S. Dumais, and D. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *SIGIR*, pages 163–170, 2008.

[54] C. Varghese and U. Shankar. Passenger vehicle occupant fatalities by day and night–a contrast. In *National Highway Traffic Safety Administration*, 2007.

[55] R. White and S. Drucker. Investigating behavioral variability in web search. In *WWW*, pages 21–30, 2007.

[56] R. White and E. Horvitz. Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM Trans. Inf. Syst.*, 27(4):23:1–23:37, 2009.

[57] R.W. White, N.P. Tatonetti, N.H. Shah, R.B. Altman, and E. Horvitz. Web-scale pharmacovigilance: listening to signals from the crowd. *JAMIA*, 20(3):404–408, 2013.