Exploring Time-Dependent Concerns about Pregnancy and Childbirth from Search Logs

Adam Fourney[1,2], Ryen W. White[2], Eric Horvitz[2]
[1]University of Waterloo, Waterloo, ON, Canada
[2]Microsoft Research, Redmond, WA, USA
afourney@cs.uwaterloo.ca, ryenw@microsoft.com, horvitz@microsoft.com

ABSTRACT
We study time-dependent patterns of information seeking about pregnancy, birth, and the first several weeks of caring for newborns via analyses of queries drawn from anonymized search engine logs. We show how we can detect and align web search behavior for a population of searchers with the natural clock of gestational physiology via proxies for ground truth based on searchers' self-report queries (e.g., [I am 30 weeks pregnant and my baby is moving a lot]). Then, we present a methodology for performing additional alignments, that are valuable for learning about the concerns, curiosities, and needs that arise over time with pregnancy and early parenting. Our findings have implications for learning about the temporal dynamics of pregnancy-related interests and concerns, and also for the design of systems that tailor their responses to point estimates of each searcher's current stage in pregnancy.

Author Keywords
Health; search; pregnancy; childbirth; temporal processes

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
Contemporary search engines often employ personalization methods, tailoring results to searchers' individual information needs. Searcher interests and preferences may be inferred from long-term query histories and interactions [4,29]. Our lives, however, are punctuated by major events such as graduation, marriage, the birth of a child, and death of a loved one. These life events often precipitate changes in our interests, preferences, and priorities [12], and the changes may evolve over the course of weeks or months (e.g., when buying a house [20]). In these periods, a person's continually evolving information needs may be poorly reflected by their prior history of search queries. However, many life events unfold along a predictable trajectory and sequence – as if following a predetermined script or schedule. If information systems can learn these schemas, and can detect when they apply, then they may better serve their users during time-dependent episodes.

CHI 2015, April 18 - 23 2015, Seoul, Republic of Korea
Copyright 2015 ACM 978-1-4503-3145-6/15/04…$15.00
http://dx.doi.org/10.1145/2702123.2702427

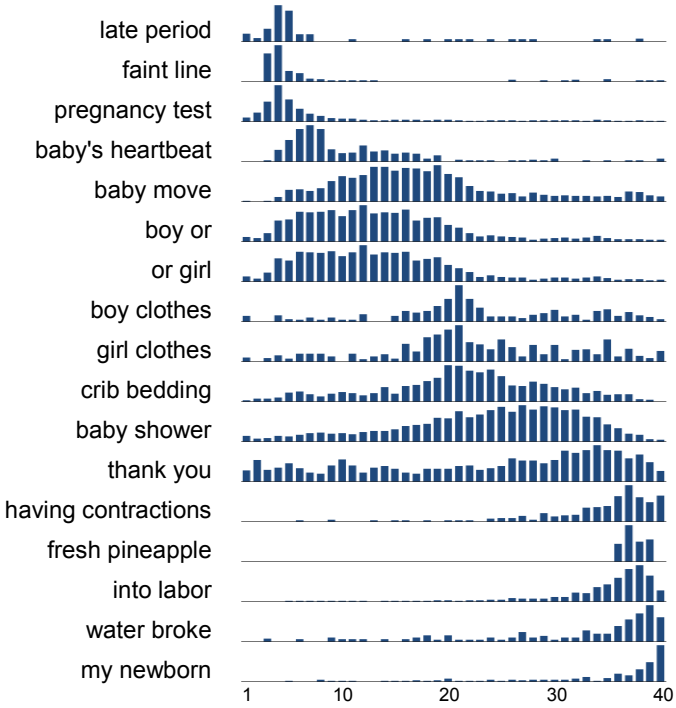

Figure 1: Histograms of query bigrams over 40 weeks of gestation. Time-dependent query volumes of each bigram (left) are displayed by gestational week for searchers who self-identified as pregnant (*n*=13,030 searchers). Each bar represents the proportion of searchers who searched at least once for bigram of interest in the corresponding gestational week. Histograms are normalized with respect to the maxima of each bigram.

We explore pregnancy and childbirth–hallmark examples of life events with well-characterized temporal dynamics. With pregnancy and childbirth, physiological processes and cultural practices (e.g., ultrasound exams, lab tests, baby showers, naming, birth announcements, etc.), conspire to stimulate curiosities, concerns, and information needs on a schedule. We demonstrate how search query logs can be leveraged to learn a detailed model of this phenomenon. Specifically,

we present methods for aligning search queries with the 40 gestational weeks of pregnancy, and show, through detailed log analyses, that information needs of new and expectant mothers wax and wane in predictable patterns over the course of the weeks of pregnancy and beyond.

Figure 1 illustrates the temporal dynamics of pregnancy-related query bigrams over time. As an example, gender prediction queries (e.g., [does carrying low mean a boy or a girl]) taper off around the 20th week of pregnancy. In the same gestational week, we observe a short-lived increase in interest for gender-specific baby clothing. The timing of these events coincides with that of the mid-pregnancy ultrasound—a screening procedure where many parents first learn of the likely gender of their child [25].

Search logs provide some more surprising signals about searcher interests. For example, Figure 1 reveals a sudden rise in searches for "*fresh pineapple*" in the final weeks of pregnancy. Pineapple contains the enzyme bromaline, which is believed to induce labor [30], but is also destroyed by food processing procedures (e.g., canning or juicing). The timing and specificity of this query activity could be interpreted as showing intent to induce labor in the final weeks of pregnancy. Per these examples, date-aligned query logs afford an unprecedented view of pregnancy and childbirth, as experienced by tens of thousands of people.

In the short-term, we envision leveraging these temporal patterns to anticipate the queries of new parents, allowing search engines to better meet the information needs of this population through methods such as personalization [4]. In the medium-to-long-term, we envision leveraging this resource to investigate public health concerns (e.g., postpartum depression, gestational diabetes, spontaneous preterm labor, pregnancy related hypertension, drug abuse in pregnancy, etc.). Also, such studies frame questions about the ethics of large-scale data analyses by bringing to light potential inferences that can be made about people—even when using anonymized logs. Thus, this work can bring to the fore important discussions about privacy.

A key insight that we leverage is that a small percentage of searchers will make unambiguous statements via their queries about the nature and timing of their experiences. Though sparse, these special queries serve as crisp entries on the timeline of a well-characterized process, such as the physiology of pregnancy and childbirth. The unambiguous entries enable a large number of other queries to be synchronized, and such data, at the population level, can be leveraged to construct more general recognizers with the ability to identify and embed more common querying behavior on the timeline.

Our primary contributions with this research are as follows:

- We present and validate a set of *self-report* query templates that can be used to identify a cohort of searchers who we strongly believe are pregnant or to have recently given birth.
- We align the query histories of these searchers to the gestational weeks of pregnancy, and we characterize how their information needs evolve over time.
- We demonstrate how the temporal characterization of pregnancy-related searches can be leveraged to detect additional pregnancies in the query logs, beyond the original set of self-reporting searchers.
- We leverage the expanded dataset to characterize differences in the information needs of first-time mothers (primipara) and those of more experienced mothers (multipara). We tailor the temporal models accordingly, affording additional opportunities for personalization.

The remainder of this paper is structured as follows: We next present related work. Then, we describe data used in our studies. We describe methods for identifying an initial set of new and expectant parents, and how due dates or delivery dates may be estimated from search histories. We characterize how searchers' information needs evolve over the gestational weeks of pregnancy. Then, we describe how these temporal patterns can be leveraged to more generally detect and classify additional searchers. Finally, we compare and contrast the information needs of primipara and multipara. We conclude with a discussion of applications, implications, and potential concerns raised by this work.

BACKGROUND AND RELATED WORK

Recent work has demonstrated the value of performing time-series analysis on large-scale search query logs. For example, Choi and Varian demonstrated how such analyses can be leveraged to estimate economic indicators such as consumer confidence or the unemployment rate [7]. Likewise, Shokouhi demonstrated how time-series analysis can be used to identify topics of periodic or seasonal interest [26]. In this paper, we are particularly interested in applications of these types of analyses to public health research. Here, Ginsberg et al., demonstrated how query data can be used to estimate the prevalence of seasonal influenza in a population [14]. White et al. developed methods to detect adverse drug interactions by monitoring a population's medication and symptom-related search queries [32,33]. West et al. leveraged query data to investigate nutrition, including exploration of a potential association between dietary sodium and rates of hospital admissions for congestive heart failure [31]. In other work, Cooper et al. characterized correlates between cancer-related search activity and: cancer incidence rates, mortality rates, as well as mentions of cancer in the media [8]. Each of these examples highlights the richness and ecological validity of query log data for characterizing phenomena relative to a public time reference (e.g., the calendar date, or the date of a news article). Our work extends this research to phenomena that have well-characterized temporal dynamics, but that are experienced at different times by different individuals. In this sense, our work more closely aligns with, and extends, Richardson's vision for long-term query log analysis [20].

Early work examining health-related web search identified pregnancy as a primary topic of interest [27]. A decade later,

pregnancy continues to top the list of health-related search topics [11,34]. In the intervening ten years, much has been written about how new and expectant parents leverage this important online resource [1,5,13,21,28]. However, query data is difficult to come by, and academics have typically relied on surveys [1], focus groups [5], or in-situ interviews [13] to learn about this phenomenon. These methods provide valuable insights into the motivations, opinions, and search strategies of expectant mothers. We complement this work by examining how pregnancy-related searches are manifest in the logs of a major contemporary search engine. This affords significant insight into how people pursue pregnancy related searches in naturalistic search settings.

Beyond logs of search queries, recent work has leveraged social media interactions to investigate aspects of pregnancy and childbirth. De Choudhury et al. characterized postpartum behavioral changes apparent in Twitter [9] and Facebook [10] postings. The work includes the construction of models that predict likely postpartum changes before the birth of the baby. Morris investigated how new mothers use social networking sites following the birth of their children [17]. Relevant to this work, Morris described typical parenting-related questions posted to these networks, and characterized how a mother's postings evolve over the months following a birth. Query log analyses serve as valuable complements to studies with social media—especially in light of recent research showing significant differences in health questions that people pose to their social networks versus in search [11].

Finally, reports appearing in the press have described how pregnancy-related searches are manifest in the query logs of major search engines [21,28]. Stephens-Davidowitz recently authored an opinion piece for the New York Times in which he compares pregnancy-related searches across 20 countries [28]. Closest to our work is a recent Google report targeted at marketers and advertisers [21]. In this work, Rost et al. leverage query log data to characterize broad trends in the information needs of new and expectant parents. Three distinct phases of parenthood are considered: pregnancy, caring for newborns, and caring for toddlers. Our research fills in many missing details by investigating the week-to-week information needs of this population.

DATA AND METHODS

We seek to characterize how information needs of new and expectant parents evolve during pregnancy. We shall first describe the general query log dataset. Then, we present a methodology for building a set of searchers who present strong evidence suggesting that they are pregnant. Finally, we describe how users' query histories are aligned with the gestational weeks of pregnancy.

Query Log Dataset

We rely exclusively on an existing proprietary dataset consisting of search queries posed to the Microsoft Bing web search engine over an 18-month period between June 2012 and December 2013. We restrict the analysis to English queries that were issued from searchers within the United States geographic locale. The resulting set consists of billions of queries, posed by tens of millions of anonymous searchers. Given the terms of use under which the data were collected, we rely only on anonymous numerical identifiers to associate queries with individual search histories.

We note that the query log dataset captures a dynamic population of searchers. Over the 18 months of data collection, new users arrived to the search engine, existing searchers departed, and anonymous user identifiers were occasionally reset. Examples of such reset events include searchers updating their Web browser or clearing their browser cookies. Users in our dataset have a median of 12 weeks of historical data.

Establishing a Set of New and Expectant Mothers

Although the query log dataset includes tens of millions of searchers, we expect only a small fraction of these individuals to be pregnant or to have had a recent childbirth during the timeframe of data collection. Limiting factors include the general fertility rate of the population, and the specific demographic of the search engine. Our study of pregnancy and childbirth requires that we identify, with high confidence, a subset of searchers whose query histories coincide with their pregnancy and childbirth experiences.

To identify searchers who may be experiencing pregnancy, we focused on individuals who had performed at least one search containing the first-person declaration of pregnancy: [I am *N* weeks pregnant]. Here, *N* is a placeholder for any integer ranging between 4 and 42 weeks. A lower limit of 4 weeks was chosen to match the time when a mother's human chorionic gonadotropic (hCG) hormone levels are sufficient to yield reliable home pregnancy test results [6]. An upper limit of 42 weeks was chosen because 97% of births occur before the 43rd week of pregnancy [16].

We removed from consideration queries containing additional qualifiers. For example, we excluded queries such as: [I think I am *N* weeks pregnant], and [when I am *N* weeks pregnant]. We considered the remaining unqualified queries as self-reports about the searchers' situations at particular times. We argue that these clear unambiguous statements are at least as reliable as survey responses, given that they are unprompted. Per these definitions, 13,030 individuals self-report as pregnant in the dataset.

Since children may be born in different gestational weeks, prenatal self-report statements are insufficient for reasoning about the timing of postpartum queries. To investigate information needs in the weeks following births, we identify searchers who self-report as having recently given birth. We focused on individuals searching with variations of the phrase [my *N* week old DEPENDANT]. Again, *N* is a placeholder for an integer, this time ranging from 1 to 24. Likewise *DEPENDANT* is a placeholder for such terms as: "son", "daughter", "newborn", "infant", "baby" or "child". Applying this methodology on our dataset yielded a set of 8,535 individuals who issued self-report queries in this way.

Alignment of Queries to Gestational Week

By design, the queries we used to construct the self-report sets each serves as a point of reference that can be used to align a searcher's entire search history with the gestational weeks of pregnancy, or with the weeks postpartum. For example, if a woman queries [I am 34 weeks pregnant and am having trouble sleeping], then we know that queries issued two weeks prior occurred in week 32 of pregnancy. Likewise, if a parent queries [how much should my 6 week old baby sleep], then we can compute the child's birth week by looking six weeks prior to the time that the query was issued.

EVOLUTION OF INFORMATION NEEDS OVER TIME

In the previous section, we described how web search queries of 13,030 self-reported expectant mothers can be aligned with the 40 weeks of gestation. Likewise, the searches of 8,535 self-reported new parents can be aligned with the weeks postpartum. In both cases, the data can be aggregated, yielding a characterization of pregnancy-related concerns unprecedented in scale. We now explore how the information needs of new and expectant parents evolve longitudinally. Potential applications include personalized search and analyses for clinical and public health.

Ebb and Flow of Pregnancy-Related Query Topics

Figure 1 detailed how interest in a set of phrases, captured as bigrams, varies over the 40 gestational weeks of pregnancy. Figure 2 presents similar information, but for queries issued over the first 24 weeks postpartum. In both figures, histogram bars correspond to weeks, and bar heights express the proportion of self-reported new and expectant mothers who issued a query containing the bigram of interest during the corresponding week. We normalize bars by dividing each entry by its histogram's maximum value.

As illustrated in Figures 1 and 2, concerns evolve throughout pregnancy, and into the weeks following the birth of a child. In early pregnancy (Figure 1), we see a continuous shift in interests, transitioning from one phase of pregnancy to the next: e.g., missing a period, taking a pregnancy test, experiencing morning sickness, detecting the baby's heartbeat on an ultrasound, and feeling the first signs of baby movement. Likewise, queries during the postpartum period (Figure 2) capture a predictable pattern of interest in newborn development, from smiling for the first time, to eating solid foods.

Beyond identifying interests linked with well-known milestones, aligning time-dependent patterns of queries across a large population may yield new insights. Figure 3 shows regularities in how expectant mothers search on pain or discomfort over 40 weeks of pregnancy. Figure 4 characterizes regularities in distributions of words about emotional states during the same timeframe. We see that queries about "*back pain*" occur more often in the first and third trimesters than in the second trimester. Likewise, occurrences of the query terms "*worried*", "*scared*" and "*terrified*" all peak at around the fifth gestational week—around the time women test positive for a pregnancy. These examples highlight the scope and personal nature of concerns expressed via queries.

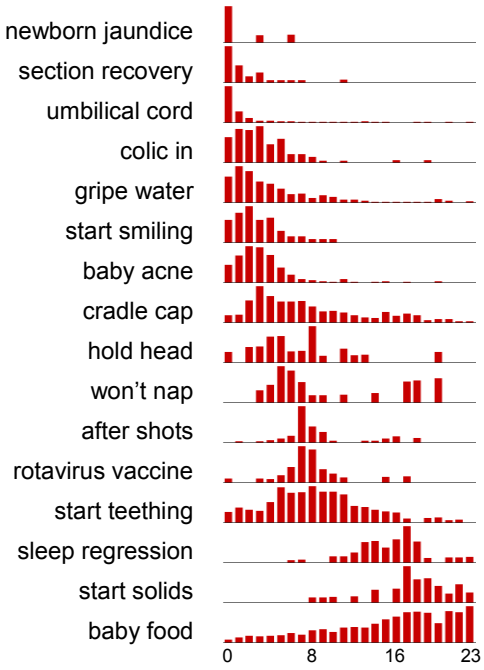

Figure 2: Histograms of query bigrams for the 24 weeks following inferred birth dates. Histograms show time-dependent query volume of bigrams for searchers who self-identified as new parents (*n*=8,535 searchers). Bars show the proportion of searchers who searched at least once for the bigram of interest within the associated week. Histograms are normalized with respect to the maxima for each bigram across the 24 weeks.

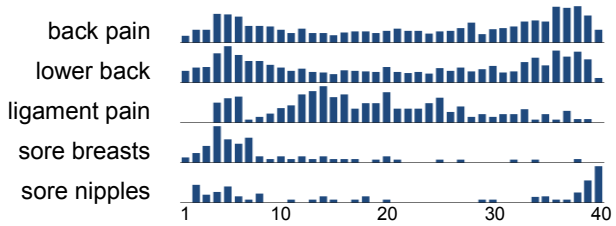

Figure 3: Histograms of pain-related bigrams over 40 gestational weeks. Columns correspond to gestational week. Histograms are normalized w.r.t. the maxima of each bigram.

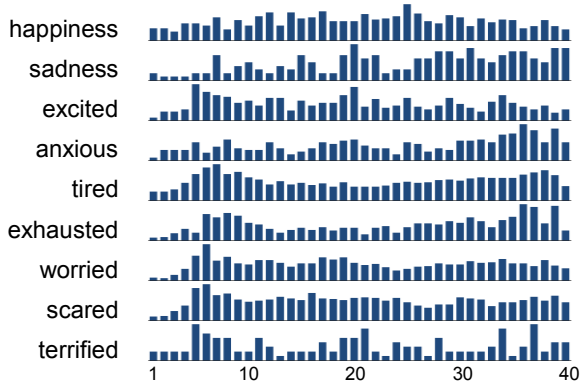

Figure 4: Histograms for terms in queries reflecting potential emotional state over the 40 gestational weeks of pregnancy. Histograms are normalized w.r.t. the maxima of each bigram.

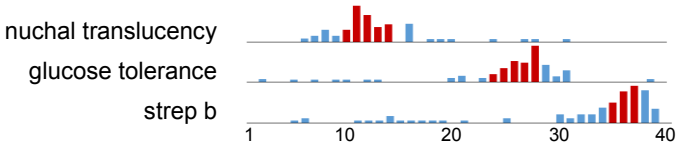

Figure 5: Histograms detail about how interest in three standard prenatal screening procedures vary over 40 gestational weeks. Bars show the proportion of searchers who have searched at least once for the bigram of interest in the associated week. Red bars report weeks in which each test is typically performed, as reported in [2], [18], and [24] respectively.

Validation of Temporal Alignment

Compelling examples of phrases exhibiting temporal regularities include bigrams associated with prenatal screening procedures (Figure 5). These screening procedures are performed at very specific times during pregnancy [2,18,24], and afford an opportunity to validate our alignment of query histories to the gestational weeks of pregnancy. For example, we note that fetal nuchal translucency is typically measured between the 10th and 14th week of pregnancy, as part of a standard screening procedure for Down syndrome [18]. In our dataset, searches including the phrase "*nuchal translucency*" align well within this range (highlighted in red in Figure 5). Likewise, interest in both the glucose tolerance test, and the test for the group B streptococcus bacteria also align well with the periods in which these tests are typically performed (weeks 24-28 in the former case [24], and weeks 35-37 in the latter [2]). These findings strongly argue for the veracity of the self-report statements, and support our query-history alignment methodology.

Implications for the Design of Personalized Systems

Temporal trends can present opportunities and challenges for personalized search [4,29]. The opportunity is afforded by recognizing that the information needs of new and expectant parents evolve along a predictable time-dependent trajectory that may be anticipated by a personalized search engine. One identified challenge is that many topics are relevant for only a short window of time during pregnancy. For personalized search engines to fully leverage such time-dependence, accurate detection of the current gestational age is required. In the next section, we present a more general methodology for both detecting pregnant searchers in the query logs and predicting gestational age.

MODEL

So far, we have characterized how information needs of new and expectant mothers evolve over the course of pregnancy. Some interests, such as those relating to medical screening procedures, arise in specific gestational weeks. In this section, we demonstrate how topics of transient interest can be leveraged to detect and align pregnancies beyond the 13,030 expectant mothers who we identified from search logs via first-person self-reporting queries.

Classification and Prediction

Mirroring our prior use of self-report templates (e.g., [I am *N* weeks pregnant]), we seek to construct a model that can: (1) classify searchers as pregnant or as not pregnant, and (2) can align their histories with the gestational weeks of pregnancy. The latter is equivalent to estimating due dates. Likewise, we would like a model that can (3) classify searchers as having recently given birth, and can (4) estimate the date of delivery when appropriate. While many solutions are possible, we achieve strong performance using a simple heuristic method employing linear regression. Our approach is motivated by a fictitious, but representative example. Suppose Alice's query history includes the following searches:

- [positive pregnancy test] on April 2, 2012
- [glucose tolerance test] on September 20, 2012
- [2 cm dilated] on November 28, 2012

With limited uncertainty, we may treat each of these queries as independent estimates of Alice's gestational age. In this case, the query [positive pregnancy test] suggests that Alice is 4 weeks pregnant on April 2nd. The query [glucose tolerance test] suggests that Alice is 28 weeks pregnant on September 20th, and so on. These estimates can be fit to a linear model. In this case, when both the dependent and independent variables are expressed in weeks, the line of best fit (via least squares) is found to exhibit a slope of 1.017. Alice appears to be gaining approximately one gestational week per week of time elapsed, thus we consider Alice's query behavior to be consistent with pregnancy. Since the slope is within a narrow threshold of acceptable values, we classify her as pregnant, and we use the regression line to align her complete query history with the inferred gestational weeks of pregnancy.

When performing linear regression on query data, we have found that ordinary least squares performs quite poorly. Inspecting errors reveals two common problems: First, searchers occasionally make typographical errors. For example, a user may type [4 weeks pregnant], and then quickly correct the query to read [24 weeks pregnant]. Second, we have observed that newly pregnant mothers often search ahead, e.g., researching later stages of pregnancy while in the first few weeks of gestation. In both cases, large residuals profoundly corrupt parameter estimates. We address this problem by adopting the Theil-Sen estimator, a non-parametric approach to robust linear regression. With the Thiel-Sen estimator, up to ~30% of the data can be corrupted, and the estimator will still provide reliable results [22].

Feature Selection and Parameter Values

When implementing this strategy we must make two choices. First, we need to decide on a threshold, $[1-\beta, 1+\beta]$ for the slope of the regression line. Second, we must select a set of query phrases to generate the pointwise estimates of the gestational week. For this, we desire phrases that (1) are associated with pregnancy, and (2) occur within a narrow range of gestational weeks. Intuitively, a good phrase will be one whose occurrences are difficult to predict prior to learning that user is pregnant, but easy to predict when we obtain

knowledge of her gestational week. This tradeoff can be expressed mathematically as information gain [23], which we estimate as reduction in entropy H, as follows:

$$IG[\![f]\!] = H[\![p(w_{year}|\, f)]\!] - H[\![p(w_{gestation}|\, f)]\!] \quad (1)$$

Here, the conditional distribution $p(w_{year}|\, f)$ characterizes how occurrences of the phrase f are distributed over the 52 weeks of a calendar year. Likewise, $p(w_{gestation}|\, f)$ characterizes how occurrences of the phrase f are distributed over the 40 weeks of gestation. In both terms, $H[p(x)]$ corresponds to the usual entropy function:

$$H[\![p(x)]\!] = -\sum_x p(x) \log p(x) \quad (2)$$

Our model selects the set of phrases from which to generate pointwise estimates by choosing the top η trigrams sorted by information gain, as defined in (1). This parameterization of phrase selection yields a model with two numeric parameters: β and η. Using a systematic parameter sweep, we explored a range of possible parameter values. We found that setting η=300, and β=0.15 achieves a good tradeoff between detecting a large number of expectant or new mothers, while achieving low average error in due date / delivery date prediction. We describe these evaluations in the next section.

Evaluation

Linear regression is used to simultaneously classify searchers as pregnant or not pregnant, and to align a searcher's query history to gestational weeks. The latter is equivalent to estimating a searcher's due date. We present an evaluation of both date prediction and classification. When working only with query log data, evaluation of model performance is challenging. The primary complication arises from the lack of ground-truth labels for searchers beyond those who unambiguously self-report their pregnancy or postpartum statuses. For this reason, we employ a range of evaluation methods, each testing a different aspect of the model.

Prediction Accuracy for Due Dates and Delivery Dates

Self-report data can be used to evaluate the accuracy of due date prediction. Evaluation is performed by learning a model from one set of self-report searchers, and then employing the model to estimate the due dates of the remaining self-report users. Results are cross-validated by repeating the procedure K=16 times, on K distinct data partitions.

Unfortunately, a complication persists even when performing cross-validation: under the above protocol, training and test data are both drawn from self-report searchers. Each of these searchers is guaranteed to have issued the query [I am N weeks pregnant]. This is an excellent feature for estimating gestational age. To mitigate this confound, we remove from consideration all queries containing self-report statements.

Taking all aforementioned factors into consideration, we found that the regression model predicts a searcher's due date to within ±0.685 weeks on average (median: 0 weeks). Predictions of delivery date based on postpartum queries (looking back in time) perform only slightly worse, achieving an average error of ±0.787 weeks (median: 0 weeks).

Estimating Classification Precision

The evaluation demonstrates that due dates and delivery dates can be accurately predicted, provided that we have information via self-reports that the searcher is pregnant, or that they have recently given birth. We also proposed applying a threshold on the slope of the regression line as a heuristic for detecting new or expectant mothers. To investigate the performance of this approach, we ran the classifier on the full 18-month query log dataset. In total, 93,463 searchers were classified as having experienced pregnancy during this timeframe, as compared to the 13,030 users in our self-report dataset. To verify the accuracy of the classification, we randomly sampled 100 of the searchers who were found using this method, omitting individuals who were part of the self-report dataset. Two researchers involved in this study then independently examined the query histories of these users, to make a determination of classification accuracy. Each rater labeled 99 searchers as having experienced pregnancy, and only one searcher who probably was not pregnant during this time. Moreover, the two raters were in perfect agreement about which searcher was falsely returned—an individual who was posing questions about their pregnant pet Rottweiler dog. We can therefore estimate that the classification heuristic achieves high precision when detecting pregnancy.

When applied to the problem of detecting mothers who had recently given birth, the classifier returned 42,548 searchers for whom delivery dates could be estimated. This represents a large increase over the 8,454 searchers appearing in the self-report dataset. The evaluation procedure was repeated. In this case, 97 searchers were labeled as having recently given birth. We again estimate that the classification heuristic achieves high precision when detecting recent childbirths. Unfortunately, recall performance cannot be easily measured in the same way. The difficulty arises from our inability to discern true negatives from false negatives without ground truth data (i.e., a person *not* searching about pregnancy may still be pregnant, and there is no way to determine this from the query logs alone).

Correlation with CDC Birth Data

As a final method to validate the model, we consider the 1,892 women whose query histories contain sufficient data to perform regression for both gestational and postpartum week alignment. In these cases, we can compute the gestational week in which each mother gave birth, by using the postpartum model to determine in which week of the year the birth took place—and then consulting the prenatal model to map this calendar week to a gestational week.

Figure 6 shows the distribution of births by gestational week, as estimated by our model (blue/solid curve). When compared to data compiled for the same span of time [16] by the Center of Disease Control (Figure 6, green/dashed curve), we find a Pearson correlation of $r = 0.97$ ($p < 0.0001$). While this measure is extremely encouraging, we note that predictions

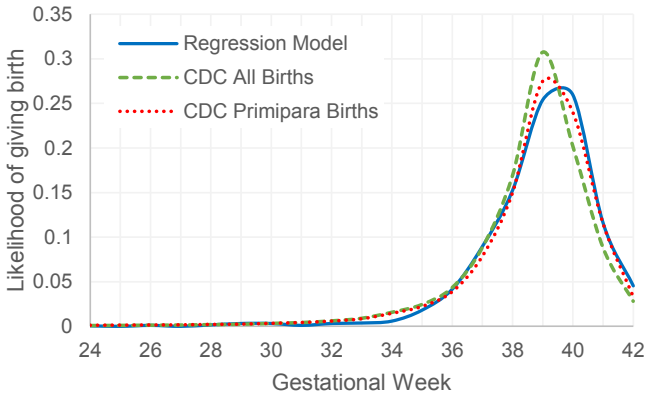

Figure 6: Distribution of birth events by gestational week (United States, 2012). The green/dashed curve represents *all* live births, as recorded in [16]. The red/dotted curve represents births to *first-time* mothers. The blue/solid curve represents the value computed by our model, using only search log data. Our model correlates well with both CDC curves (Pearson $r = 0.97$, $p < 0.0001$ for all births, and $r = 0.995$, $p < 0.0001$ for births to first-time mothers).

made by our model skew slightly toward later gestational weeks. There are a number of possible explanations for this skew. For example, we cannot rule out systematic noise in the data. Systematic noise could arise from the already noted tendency for some searchers to search ahead in early pregnancy. It is possible that our regression technique does not completely correct for this phenomenon.

The skew towards later weeks could also arise if the user population is biased towards *primipara* (first-time mothers), as primipara tend to give birth later in pregnancy than *multipara* (mothers with at least one prior live birth) [16]. This explanation is compelling: when the CDC data is filtered to include only first-time mothers (Figure 6, red/dotted curve), we find a Pearson correlation of $r = 0.995$ ($p < 0.0001$). A bias towards primipara could arise if the query logs skew towards a younger population with people who have had fewer childbearing years. Alternatively, first-time mothers may issue more queries, and are thus more likely to be detected by our model. This latter hypothesis motivates a more detailed comparison of primipara and multipara in the search logs. We explore this topic in the next section.

INFORMATION NEEDS OF PRIMIPARA VS. MULTIPARA

We now explore differences in the information needs of primipara and multipara births. We conjecture that these two groups have different needs and interests when searching for information related to pregnancy.

Establishing Primiparous and Multiparous Searchers

To identify primipara in the query logs, one author sampled 2000 queries containing the phrase "*first pregnancy*", then identified queries that suggest a searcher's primiparous status. For example, the query [length of labor first pregnancy] meets this criteria while the query [first pregnancy appointment] does not. In total, 1180 queries were found to unambiguously identify 1678 searchers as first-time mothers (queries may be searched by more than one individual). To assess the reliability of the inclusion criteria, a second researcher examined 200 of the 2000 sampled queries. A high inter-rater reliability was observed between the two judges (Cohen's Kappa of $\kappa = 0.9389$).

Top Primipara Bigrams	Top Multipara Bigrams
baby shower	for kids
baby clothes	for toddlers
baby registry	year old
baby cribs	pregnancy showing
baby furniture	birthday party
nursery ideas	2 year
shower invitations	in children
baby nursery	year olds
stretch marks	in toddlers
birthing classes	double stroller

Table 1: Top 20 bigrams (excluding bigrams stating live birth order) for distinguishing primipara from multipara.

The procedure was repeated to identify multipara, this time sampling queries containing the phrases "*Nth pregnancy*", or "*previous pregnancy*" (where *N* is any integer greater than 1). In total, 1594 queries were found to unambiguously identify 1709 searchers as multipara. Again, a second researcher examined 200 of the 2000 queries, and again high inter-rater reliably was attained (Cohen's Kappa of $\kappa = 0.9386$).

Distinctiveness of Needs of Primipara and Multipara

As a first step we identified information needs discriminating primipara and multipara. To do this, we examined larger search histories of primiparous and multiparous searchers, and computed normalized information gain [23] for each bigram or trigram encountered in their respective query streams.

Table 1 presents the top 10 most discriminative bigrams for each class, excluding bigrams that explicitly mention live birth order (e.g., "*first pregnancy*", "*second pregnancy*", etc.). The resultant features appeal to our intuition, and strongly suggest that our sets of primipara and multipara are indeed capturing real phenomena. Primipara are more likely than multipara to: plan baby showers, decorate nurseries, attend childbirth classes and worry about stretch marks. Conversely, multipara are more likely to search about: toddlers, birthday parties, and double strollers.

Time-Dependent Differences in Information Needs

Having identified small populations of multipara and primipara in the query data, it is possible to compare and contrast the evolution of information needs for the groups. Figure 7 presents how a set of trigrams are manifest in both populations. As before, the horizontal axis denotes the gestational week of pregnancy and the vertical axis denotes the proportion of the population issuing at least one query containing

the trigram of interest in a given week. In the first three examples, the differences between primipara and multipara are chiefly ones of magnitude: assumed first-time mothers are more likely to ask which foods to eat when pregnant, or to ask if a particular experience is normal, and multipara are more likely to search for a "*due date calculator*". We hypothesize that multipara are more likely to be familiar with the terminology generally used to describe these tools, or to even know that such tools exist online.

The three examples at the bottom of Figure 7 characterize query phrases where the differences between primipara and multipara are chiefly those of timing. For example, while both groups appear to begin asking when will they "*start to show*" nearly immediately after learning of pregnancy, multipara stop asking approximately four weeks before primipara. This agrees with the belief that multipara will begin to show earlier in pregnancy than primipara [35]. Likewise, occurrences of the phrase "*feel the baby*" rise and fall earlier for multipara than for primipara. This agrees with the observation that quickening (first perception of fetal movement) occurs several weeks earlier for multipara than for primipara [15]. These examples should not be misinterpreted as confirmations of these presumed phenomena; that these are beliefs are popular is a significant confound, and may themselves explain patterns in observed querying behavior.

Finally, query activity including the trigram "*signs of labor*" is nearly identical for primipara and multipara. However, the curve for multipara is slightly shifted towards earlier gestational weeks, agreeing with the observed phenomenon that multipara give birth slightly earlier than primipara [16].

These examples demonstrate that, in the case of pregnancy, temporal models can be refined and tailored to reflect differences between subpopulations of users. These refined models may afford additional or improved opportunities for search personalization, and may yield additional insight when leveraging the query logs to pursue other questions of scientific interest.

APPLICATIONS

A central thesis of this research is that a new or expectant mother's information needs depend on many factors including her gestational week at time of query, and her previous experience with childbirth. In many cases, these details can be inferred from a searcher's query history. This affords the possibility of constructing information interfaces that tailor results to the searcher's specific situation. We outline some of these possibilities below, then describe how these models could be applied to investigate topics of public health.

Supporting Mothers through Personalized Search

For example, Figure 3 reports that expectant mothers issue many queries with the term "*exhausted*" in both early and late pregnancy. Fatigue in early pregnancy has been linked to hormonal changes, while fatigue in late pregnancy has been linked to iron deficiencies and trouble sleeping [3]. Search engines could leverage temporal information about these searches to tailor relevance rankings of fatigue-related pages in cases where the searcher's stage of pregnancy is known.

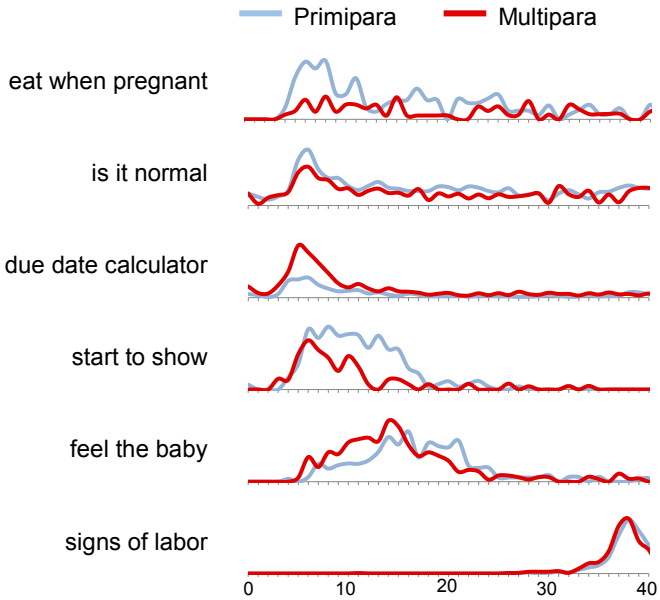

Figure 7: Graphs characterizing the differences between the search behavior of primipara (blue) and multipara (red) for a range of trigrams. Curves report temporal changes in the proportion of primipara or multipara searching with the trigram over the gestational weeks of pregnancy. The top three graphs exemplify differences in query volume. The bottom three graphs exemplify differences of timing.

Our research also raises the possibility for personalized search to take a more direct role in advising new and expectant mothers. For example, if an expectant mother searches for airline tickets in a vertical search engine, the flights returned could each include a brief description of the airline's policy on flying while pregnant. This is particularly important during the later stages of pregnancy when there may be more restrictions over whether expectant mothers can fly. Likewise, if a new mother searches for information about high fevers in children, and it is known that her child is less than two months old, the search engine could advise the searcher of a potential medical emergency and the need to seek professional medical attention.

Supporting New and Expectant Mothers through Aggregation of Pregnancy Experiences

Online pregnancy resources often serve as community support [13] to reassure expectant mothers that physical discomfort (e.g., morning sickness, soreness, back pain), and emotional stressors (e.g., worry and fatigue), are temporary experiences. Query logs provide quantifiable data to support these assertions, and could serve a role in supporting mothers in times of discomfort or stress. For example, women searching for information about back pain in early pregnancy might be comforted to know that: (1) many expectant mothers have issued similar queries at the same stage of pregnancy, and (2) the queries do not tend to persist into the second trimester.

Applications for Public Health Research

In the medium-to-long-term, we envision leveraging this resource to investigate matters of public health concern (e.g., postpartum depression, gestational diabetes, spontaneous preterm labor, drug abuse in pregnancy, neonatal jaundice, etc.) As an example, we consider the topic of self-induced labor. Figure 8 characterizes when different herbal or homeopathic methods of labor induction are most commonly searched. Interest in these methods experiences an uptick in week 36 and peaks in weeks 37 or 38. Whether such methods work as may be believed, this finding is troubling: medical practitioners caution against inducing labor prior to week 39 of pregnancy [19]. The primary concern with earlier induction of labor is that, while a child born in week 37 is not premature, there may be dating inaccuracies. Consequently, early induction raises the risk of complications resembling those experienced with premature births. Query logs alone may not be sufficient to determine if women are following through with their intention to induce labor, or if common methods are effective. Nevertheless, this could justify further studies investigating the impact of this phenomenon.

ETHICS AND PRIVACY

All data access and analysis performed for this research was done in accordance with Bing's published end-user license agreement, which specifies that user data may be used for research purposes and to improve the search experience. Our work was conducted offline, on data collected to support existing business operations, and did not influence the presentation of search results or other aspects of the user experience. All data were anonymized prior to analyses. The Ethics Advisory Committee at Microsoft Research considers these precautions sufficient for triggering the Common Rule, exempting this work from detailed ethics review.

Nevertheless, important ethical questions come to the fore with research that shows how inferences can be made about people based on their activities with online services. We realize that many may view pregnancy and childbirth as sensitive subjects. The results reported in this paper are aggregates of many searchers, and characterize positive or unexceptional aspects of experiences with pregnancy and newborns. However, search logs contain a broad spectrum of pregnancy-related topics. They capture the concerns of those who are struggling to conceive a child and those who have experienced a loss. They include questions about paternity and about abortion. And, they provide evidence of behaviors that may place children at risk (e.g., abuse of recreational drugs while pregnant). Studying these and other sensitive phenomena can advance research in clinical medicine, public health, psychology, and sociology. However, we need to reflect with the broader research community and with the public at large about the ethical and privacy implications of the research. We hope that studies like this one will help to frame the discussion about acceptable or expected uses of large-scale behavioral data when seeking deeper insights about human behavior, whether the data is acquired via search services or via publicly-visible posts, such as those made to Twitter [9].

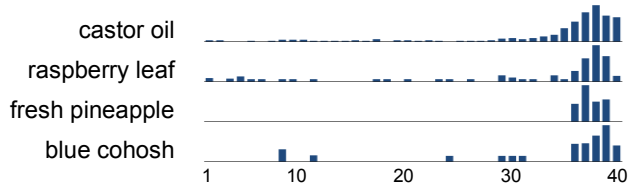

Figure 8: Histograms characterizing how the interest in self-induction of labor, varies over the 40 gestational weeks of pregnancy. Activity begins to rise in week 36 and peaks around week 38. Induction before week 39 is not recommended.

DISCUSSION AND CONCLUSIONS

Pregnancy and childbirth have profound impact on the lives of expectant and new mothers, and information seeking leaves a clear signature in logs of their search activity. We presented methods for characterizing these dynamic processes via retrospective analyses of logged search engine queries. We also performed additional alignments to refine our understanding of these phenomena.

The analysis is limited in several ways. Our method considers the appearance of special first-person queries as strong evidence of pregnancy. However, we could not contact participants directly. Thus, we had no way to verify that they were in fact pregnant. Follow-on studies with new and expectant mothers are required to validate our approach. Our study also focuses only on search activity in the United States. Pregnancy is treated differently in other countries [28] and further analysis of related search activity in different geographic regions is needed to understand the generalizability of our methods to different countries and cultures.

Pregnancy is an important topic and our study demonstrates the value of logs as a broad lens on relevant interests and concerns. Our findings have implications for learning about the temporal dynamics of pregnancy-related interests and concerns, and also for the design of systems that tailor their responses to point estimates of searchers' current stage in pregnancy. While our work focuses on pregnancy and childbirth, the methods presented in this paper will generalize to other life events, such as the treatment of a chronic disease, where well-characterized temporal dynamics of physiology or other processes can be synchronized with search queries.

REFERENCES

1. Allen, K. Parents online. *Pew research center's internet & American life project*, 2002. http://www.pewinternet.org/2002/11/17/parents-online/.

2. American Congress of Obstetricians and Gynecologists. Group B streptococcus and pregnancy. 2011. http://www.acog.org/Patients/FAQs/Group-B-Streptococcus-and-Pregnancy.

3. Amy, F. First trimester fatigue. *University of Rochester medical center: health encyclopedia*, 2014. http://www.urmc.rochester.edu/encyclopedia/content.aspx?ContentTypeID=134&ContentID=4.

4. Bennett, P.N., White, R.W., Chu, W., et al. Modeling the impact of short- and long-term behavior on search personalization. In *Proc. SIGIR '12*, ACM (2012), 185–194.

5. Bernhardt, J.M. and Felter, E.M. Online pediatric information seeking among mothers of young children: results from a qualitative study using focus groups. *Journal of Medical Internet Research 6*, 1 (2004).

6. Butler, S.A., Khanlian, S.A., and Cole, L.A. Detection of early pregnancy forms of human chorionic gonadotropin by home pregnancy test devices. *Clinical Chemistry 47*, 12 (2001), 2131–2136.

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

8. Cooper, P.C., Mallon, P.K., Leadbetter, S., et al. Cancer internet search activity on a major search engine, United States 2001-2003. *J Med Internet Res 7*, 3 (2005), e36.

9. De Choudhury, M., Counts, S., and Horvitz, E. Predicting postpartum changes in emotion and behavior via social media. In *Proc. CHI '13*, ACM (2013), 3267–3276.

10. De Choudhury, M., Counts, S., Horvitz, E.J., and Hoff, A. Characterizing and predicting postpartum depression from shared Facebook data. In *Proc. CSCW '14*, ACM (2014), 626–638.

11. De Choudhury, M., Morris, M.R., and White, R.W. Seeking and sharing health information online: comparing search engines and social media. In *Proc. CHI '14*, ACM (2014), 1365–1376.

12. Duhigg, C. How companies learn your secrets. *The New York Times*, 2012. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

13. Gibson, L. and Hanson, V.L. Digital motherhood: how does technology help new mothers? In *Proc. CHI '13*, ACM (2013), 313–322.

14. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature 457*, (2009), 1012–1014.

15. Levene, M.I., Tudehope, D.I., and Thearle, M.J. Definitions and terminology. In *Essentials of Neonatal Medicine*. Wiley, 2000, 9.

16. Martin, J.A., Hamilton, B.E., Osterman, M.J., Curtin, S.C., and Mathews, T. Births: final data for 2012. *National Vital Statistics Report 62*, 9 (2013), 1–87.

17. Morris, M.R. Social networking site use by mothers of young children. In *Proc. CSCW '14*, ACM (2014), 1272–1282.

18. Nicolaides, K.H., Azar, G., Byrne, D., Mansur, C., and Marks, K. Fetal nuchal translucency: ultrasound screening for chromosomal defects in first trimester of pregnancy. *BMJ : British Medical Journal 304*, 6831 (1992), 867–869.

19. Oshiro, B.T., Henry, E., Wilson, J., Branch, D.W., and Varner, M.W. Decreasing elective deliveries before 39 weeks of gestation in an integrated health care system: *Obstetrics & Gynecology 113*, 4 (2009), 804–811.

20. Richardson, M. Learning about the world through long-term query logs. *ACM Trans. Web 2*, 4 (2008).

21. Rost, J., Johnsmeyer, B., and Mooney, A. Diapers to diplomas: what's on the minds of new parents. *Think with Google*, 2014. http://www.thinkwithgoogle.com/articles/new-parents.html.

22. Rousseeuw, P.J. and Leroy, A.M. Other techniques for simple regression. In *Robust Regression and Outlier Detection*. John Wiley & Sons, 2003, 67–78.

23. Russell, S. and Norvig, P. Choosing attribute tests. In *Artificial Intelligence a Modern Approach*. Prentice Hall, 2003, 660.

24. Sacks, D.A., Greenspoon, J.S., Abu-Fadil, S., Henry, H.M., Wolde-Tsadik, G., and Yao, J.F.F. Toward universal criteria for gestational diabetes: The 75-gram glucose tolerance test in pregnancy. *American Journal of Obstetrics and Gynecology 172*, 2, Part 1 (1995), 607–614.

25. Salomon, L.J., Alfirevic, Z., Berghella, V., et al. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound in Obstetrics & Gynecology 37*, 1 (2011), 116–126.

26. Shokouhi, M., and Radinsky, K. Time-sensitive query autocompletion. In *Proc. SIGIR '12*, ACM (2012), 601–610.

27. Spink, A., Yang, Y., Jansen, J., et al. A study of medical and health queries to web search engines. *Health Information & Libraries Journal 21*, 1 (2004), 44–51.

28. Stephens-davidowitz, S. What do pregnant women want? *The New York Times*, 2014. http://www.nytimes.com/2014/05/18/opinion/sunday/what-do-pregnant-women-want.html.

29. Teevan, J., Dumais, S.T., and Horvitz, E. Personalizing search via automated analysis of interests and activities. In *Proc. SIGIR '05*, ACM (2005), 449–456.

30. Tiran, D. Complementary therapies: your questions answered. Use of pineapple for induction of labour. *Pract Midwife 12*, 9 (2009), 33–4.

31. West, R., White, R.W., and Horvitz, E. From cookies to cooks: insights on dietary patterns via analysis of web usage logs. In *Proc. WWW '13* (2013), 1399–1410.

32. White, R.W., Harpaz, R., Shah, N.H., DuMouchel, W., and Horvitz, E. Toward enhanced pharmacovigilance using patient-generated data on the Internet. *Clinical Pharmacology & Therapeutics 96*, 2 (2014), 239–246.

33. White, R.W., Tatonetti, N.P., Shah, N.H., Altman, R.B., and Horvitz, E. Web-scale pharmacovigilance: listening to signals from the crowd. *Journal of the American Medical Informatics Association, 20*, 3 (2013), 404–408.

34. Google's top health searches for 2012: cancer, pregnancy symptoms and ... hemorrhoid? *MedCity news*. http://medcitynews.com/2012/12/googles-top-health-searches-for-2012-cancer-pregnancy-symptoms-and-hemorrhoid/.

35. Pregnant again: What to expect this time around. *BabyCenter*. http://www.babycenter.com/0_pregnant-again-what-to-expect-this-time-around_10305185.bc