# Improving Search Engines via Large-Scale Physiological Sensing

Ryen W. White
Microsoft Cortana
Bellevue, WA 98004
ryenw@microsoft.com

Ryan Ma
Microsoft Bing
Beijing, China 100080
ryanma@microsoft.com

## ABSTRACT

Result ranking in commercial web search engines is based on a wide array of signals, from keywords appearing on web pages to behavioral (clickthrough) data aggregated across many users or from the current user only. The recent emergence of wearable devices has enabled the collection of physiological data such as heart rate, skin temperature, and galvanic skin response at a population scale. These data are useful for many public health tasks, but they may also provide novel clues about people's interests and intentions as they engage in online activities. In this paper, we focus on heart rate and show that there are strong relationships between heart rate and various measures of user interest in a search result. We integrate features of heart rate, including heart rate dynamics, as additional attributes in a competitive machine-learned web search ranking algorithm. We show that we can obtain significant relevance improvements from this physiological sensing that vary depending on the search topic.

## CCS CONCEPTS

•**Information systems** →*Data stream mining; Learning to rank;*

## 1 BACKGROUND AND INTRODUCTION

Consumer fitness wearables such as FitBit and Apple Watch can track signals such as sleep and physical activity from large populations of users. There has been significant research on the application of such wearable devices to persuade people to live healthier lives [9] or recognize their physical activities [14]. Search logs have been used for tasks ranging from understanding search behaviors [17] to improving ranking [3, 11]. Dwell time estimates on landing pages provide insight into people's engagement [7] although time alone is insufficient to determine relevance [12].

Recent work has characterized relevance using neurological and physiological methods [5, 8, 15, 16]. Unlike neurological measures, physiological signals can be collected fairly easily. These have been used for applications such as search personalization [1] and satisfaction modeling [6]. One sensor that is common to many wearable devices is a heart rate (HR) monitor, often implemented as an optical sensor. Wearable devices record the HR in different ways, depending on the nature of the sensor and their battery utilization strategies. By joining HR signals with search engine activity data, new opportunities emerge to understand relevance and improve search engine effectiveness. We conjectured that by using HR data from many users of wearable devices, and tracking how their HR changes as they search the web, we can better estimate the relevance of specific pages. Figure 1 shows a fictitious, but realistic, example of these temporal dynamics. Initially the HR of the user is steady, but there is a step change in the HR once they visit the search result (landing page). Characteristics of HR dynamics (e.g., the delta in the HR between the search engine result page (SERP) and the landing page) may offer insight on landing-page relevance for the current searcher, and future searchers when aggregated across queries and integrated into a generic ranking algorithm.

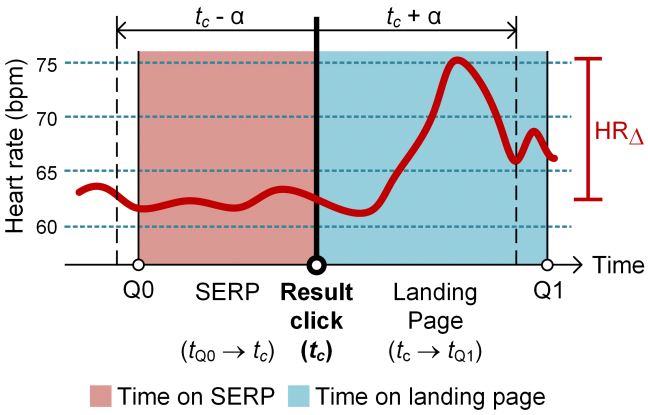


**Figure 1: HR signal from a wearable device over the course of a typical search engine interaction comprising query submission ($Q0$), result page examination, result clickthrough, examination of a landing page, and a follow-on query ($Q1$). The HR (red line) is observed to spike once the user views the landing page. Features of the change in HR (e.g., the magnitude of the change in HR from SERP to landing page – denoted $HR_\Delta$) can be used as relevance signals. There is an observation period before and after the click at time $t_c$ during which feature values can be calculated in offline settings.**

Our research extends previous work on the use of physiological feedback, which has traditionally been used for personalization [8, 13, 15, 16], limiting its applicability to instrumented users only. In this paper, we show that we can leverage physiological signals *at scale* to improve the performance of search engine ranking algorithms. We show that we can learn ranking models from a subset of the population for whom we collect physiological signals that can significantly improve search relevance for all search engine users, even the large subset of searchers who do not use wearables.

The primary contributions of this paper are:


*SIGIRfi17, August 7–11, 2017, Shinjuku, Tokyo, Japan*

DOI: http://dx.doi.org/10.1145/3077136.3080669

**Table 1: Statistics for the dataset described in Section 2. Statistics are shown for the time period from December 2015 – June 2016. The length of the observation period ($\alpha$) is 300 seconds. *has history* denotes presence of HR in the user history from September 2015 – November 2015 (inclusive).**

| Dataset Statistics | Statistic |
|---|---|
| Search and HR data joined log duration | 7 months |
| # users with HR data connected to $\geqslant$ 1 query | 23,319 |
| # queries in total | 26,327,967 |
| # result clicks in total | 27,524,934 |
| # unique query-URL pairs | 2,559,857 |
| # heart rate measurements | 28,850,947,740 |
| # heart rate measurements with quality $\geqslant$ 5 | 19,924,131,989 |
| **Where $\alpha$ = 300s:** | |
| # clicks with HR data before | 832,736 |
| # clicks with HR data after | 3,063,499 |
| # clicks with HR data after, has history | 2,219,971 |
| # clicks with HR data before, after, has history | 639,346 |

- Present the first study on using *large-scale* physiological signals to train generic ranking algorithms that yield relevance gains;
- Demonstrate the clear relationship between HR signals and more traditional indicators of searcher interests or relevance, namely human relevance judgments, result clickthrough rate, and dwell time on landing pages, and;
- Understand the impact of query topic on the utility of physiological signals for improving search relevance.

# 2 METHODS

## 2.1 Data

We use the data from the Microsoft Band and the Bing search engine. Data were collected from a sample of 23k users of the Band device who consented to connect their search and wearable accounts through a common user identifier. This consent was collected to generate richer insights about people's sleep and activity by capitalizing on additional data from the search engine. To reduce the impact of cultural and linguistic factors, participants were drawn from the United States geographic locale. Seven months of query-click logs from Bing combined with Band data (both from December 2015 to June 2016 inclusive) for these users were used to generate features of search activity and physiology. For a subset of the users in our dataset for whom we had least 1000 HR readings in the three months from September to November 2015 inclusive we could establish a normative HR, enabling the generation of additional features related to deviations from their normal HRs.

The Bing search logs contained millions of search engine interactions comprising anonymized user identifiers, user actions (queries and search-result clicks), and timestamps for those actions. The Band records HR data in addition to other physiology, sleep, and physical activity signals. To conserve battery life, the Band records continuous HR data once per second for a sampling duration of one minute out of every 10 minutes, unless the user is engaged in workout activity (where logging is continual), and two minutes out of every 10 minutes when asleep. That means that to obtain a reading on the HR in the time period before or after the click (i.e., $t_c \pm \alpha$ in Figure 1) we need to focus on the specific clicks for which tracking is available (i.e., approximately 10% of the queries).

HR is tracked on the Band using an optical sensor sampling R wave-to-R wave (RR) intervals at a rate of 60Hz. The resultant HR estimate is logged once per second in beats per minute. The HR estimate is accompanied by a quality reading from the device between 0 and 10 (higher is better). We only use readings with a rating of five or greater in our analysis. Logs were uploaded to a remote server when users synced their Band with their smartphone.

## 2.2 Combining Data Streams

The two datasets described in the previous section were joined together based on a common user identifier and timestamp, allowing the HR data to be associated directly with search events given searcher consent. Being able to associate search events with physiological signals is critical for feature generation and the envisaged application of these data in search result ranking. Since the focus was on being able to infer relevance of landing pages for a given query, we centered the analysis on those pages. We tracked HR in a time window extending $\alpha$ seconds before and after the time of the click ($t_c$). We experimented with $\alpha \in \{30s, 120s, 300s\}$). We focus on $\alpha = 300s$ since it allows us to maxmize query coverage. The tracking period was terminated if the observation window expired or there was another search action (query or click) before the end of the observation period. Examples of both of these scenarios are provided in Figure 1: early termination of $t_c - \alpha$ at $Q0$ and completion of $t_c + \alpha$ just prior to $Q1$. The HR readings during these periods were recorded to compute descriptive statistics that could be used as features in a machine-learned search-result ranking model.

Basic dataset statistics are shown in Table 1 for $\alpha = 300s$, including the number of clicks for which we have HR data at different points in time with respect to the click. One reason that there were fewer clicks with HR pre-click than post-click is that the time until the previous event was significantly lower pre-click (median SERP dwell time=30s) than the time until the next event post-click (median landing page dwell time=160s).

## 2.3 Featurization

To use the HR signals in ranking, we had to first create features for each instance of a query-click pair in our log data and then aggregate those features across all instances of the query-click pair in our data. To improve the reliability of the aggregation, we hash the URL to normalize case, remove trailing slashes, and collapse HTTP and HTTPS protocols. We computed descriptive statistics (average, standard deviation, minimum, maximum, range) on the HR signals before and after the click separately, over the time period extending from the click to the next/previous action or $\alpha$, whatever comes sooner. Changes in these feature values from before the click to after the click ($HR_\Delta$) were also recorded. The average and the standard deviation for each of these features across all instances of the query-click pair were recorded and used as features. Features were aggregated per URL, per site, per query-URL pair, and per query-site pair. The output of each of these steps was a query and/or URL/site pair with an associated set of around 20 features grouped into the following classes: (i) *SERP*: Descriptive features of HR

**Table 2: Correlations between the feature classes for query-URL pairs and their (i) average dwell time (Pearson correlation), (ii) relevance judgments (Spearman correlation), and (iii) clickthrough rate (CTR) (Pearson correlation). Maximal values within each of the columns are underlined. All correlations are significant at $p < 0.01$.**

| Feature class | Dwell time | Relevance | CTR |
|---|---|---|---|
| SERP | 0.1843 | 0.3699 | <u>0.4862</u> |
| LandingPage | 0.1683 | 0.1372 | 0.4074 |
| Transition | <u>0.2648</u> | <u>0.3708</u> | 0.4513 |
| History | 0.1622 | 0.3290 | 0.3556 |

when the user is on the SERP before they click on a result (27.19% coverage of the query-URL pairs with one or more HR features); (ii) *LandingPage*: Descriptive features of HR signals on the landing page (100% coverage); (iii) *History*: Variations from normal HR for user, established from the three months of user history (Sep–Nov 2015) (72.47% coverage), and (iv) *Transition*: Differences between HR on the SERP and HR on the landing page (27.19% coverage).

## 2.4 Data Analysis

To understand the associations between physiological changes and searcher interests, we computed the correlations between each of four feature classes and three popular proxies for interest or relevance: (i) *Dwell Time*: Median duration in seconds on the landing page for this query aggregated over many users, computed based on time from click to the next recorded action; (ii) *Relevance Judgments*: A rating of the landing page relevance for the query provided by trained human judges as part the standard Bing relevance assessment process. Relevance ratings were provided on a five-point scale: Bad, Fair, Good, Excellent, and Perfect, and; (iii) *Clickthrough Rate (CTR)*: The rate with which the landing page is clicked for the query when returned in its search results, based on a heldout set of 12 months of logs from September 2014 to August 2015.

Table 2 reports the correlations between each feature class and the measures of interest or relevance. We pooled all features in each class into a single list and compared their values with the measures. HR features are fairly well correlated with each of the three measures. The *SERP* and *Transition* feature classes are particularly well correlated. One explanation is that *SERP* might capture some of the interest in the landing page before the result click occurs, e.g., information residing in the result caption. The *Transition* feature class could reflect the impact of the click on searcher physiology.

# 3 RANKING USING HEART RATE

## 3.1 Models and Experimental Setup

To learn effective rankings and to explore the importance of features related to HR, we use the LambdaMART algorithm [18]. LambdaMART is a state-of-the-art ranking algorithm based on boosted regression trees. Compared with other ranking approaches, it is typically more robust to sets of features with widely varying ranges of values, such as categorical features. Since LambdaMART produces a tree-based model, it can be used as a feature selection algorithm or to rank features by their importance (Section 3.2.1).

**Table 3: Average percentage NDCG change over baseline for each feature aggregation strategy. Yellow denotes little or no change and green denotes large change. * denotes differences significant at the top-ranked position at $p < 0.05$ using $t$-tests. Coverage denotes the percentage of the test set with at least one judged URL with at least one HR feature.**

| Metric | URL* | Site | Query-URL* | Query-Site* |
|---|---|---|---|---|
| NDCG@1 | 1.51% | 0.54% | 6.99% | 3.33% |
| NDCG@2 | 1.37% | 0.95% | 4.83% | 4.15% |
| NDCG@3 | 1.27% | 0.73% | 3.94% | 3.97% |
| NDCG@4 | 0.95% | 0.60% | 3.33% | 3.79% |
| NDCG@5 | 1.05% | 0.66% | 2.77% | 3.70% |
| Coverage | 47.06% | 98.74% | 16.42% | 17.40% |

As a baseline we use LambdaMART (with 500 decision trees and learning rate of 0.1) on over 500 features including page and query content, hyperlink structure, and aggregated search activity. This gave us a competitive baseline against which to measure the impact of HR features. We train on a set of over 500k queries (and 10M query-URL pairs), validate on 60k queries, and test on a set of 16k queries – with no overlap between the query sets. The queries in the datasets were sampled from Bing logs. Query-URL pairs were labeled with five-point relevance judgments from trained human judges as mentioned earlier. We added HR-based features to the baseline ranker, retrain, and measure the relevance of the top-ranked results. We experimented with adding features for each aggregation strategy (URL, query-URL, etc.) separately and report on the performance relative to the baseline.

## 3.2 Experimental Results

*3.2.1 Overall Performance.* We now present the results across all 16k queries in our test set. We use normalized discounted cumulative gain (NDCG) to measure relevance [10] at each of the top five rank positions. Table 3 presents the percentage gain in NDCG at each of the top-five rank positions using each feature aggregation strategy. Absolute NDCG numbers are proprietary. The results show that performance improves as the aggregation strategy becomes more specific, e.g., query-URL generates the most significant gains within the covered query segment, while still covering a significant fraction of the test queries (16.4%). Inspecting the evidential weights in the learned model we find that two classes are especially important: *SERP* and *Transition* (matching the trends in Table 2) with additional contribution from landing page features and deviation from the normal HR for the current user.

*3.2.2 Topic Effects.* Physiological changes can distinguish emotive responses to certain stimuli [4]. We therefore wanted to understand whether topic affected the utility of HR features in ranking. To do this, we obtained topic classifications for around 16% of the queries in the test set by joining them with a separate set of Bing logs, where queries were already classified into top-level Open Directory Project (ODP, dmoz.org) categories (e.g., Health, Sports) using methods similar to [2]. Table 4 presents the ranking results, in terms of the percentage change in NDCG over the baseline at

**Table 4: Average percentage NDCG change over baseline (general model applied to queries in each category) for each of the top-level ODP categories. Colors denote the size and directionality of the change (darker green = more positive, darker red = more negative, yellow = no change). Also shown are the number of queries within our test set with each category label and the percentage of queries in that subset for which signals from HR data lead to any changes in retrieval performance versus the baseline. * denotes categories with significant changes at $p < 0.05$ at the top-ranked position using $t$-tests. "Regional" and "World" ODP categories were excluded since they are location-based and typically unrelated to search interests.**

| Category | Business* | Arts* | Computers* | Reference | Society | Shopping | Recreation* | Sports* | Science* | Home* | Games* | Health | News | Kids & Teens | Adult |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NDCG@1 | 3.91% | 7.94% | 9.12% | 3.82% | 2.18% | 4.21% | 12.37% | 18.99% | 9.87% | 14.89% | 13.15% | -8.08% | -6.46% | 19.81% | 10.93% |
| NDCG@2 | 2.84% | 7.66% | 5.79% | 1.24% | 0.19% | 3.29% | 8.41% | 6.07% | 3.27% | 11.24% | 11.12% | -2.44% | -0.14% | 13.88% | 8.83% |
| NDCG@3 | 3.14% | 5.62% | 4.96% | 2.07% | 1.91% | 2.17% | 6.10% | 4.97% | 1.33% | 4.55% | 6.93% | -1.47% | 1.23% | 7.27% | 14.12% |
| NDCG@4 | 1.78% | 4.40% | 4.40% | 2.55% | 2.28% | 3.12% | 4.82% | 4.27% | 2.97% | 4.06% | 5.18% | -0.87% | -1.69% | 7.60% | 12.56% |
| NDCG@5 | 1.51% | 3.74% | 3.82% | 2.68% | 0.99% | 2.52% | 3.81% | 4.17% | 2.43% | 3.68% | 5.54% | 0.20% | -5.17% | 6.59% | 7.24% |
| # queries | 570 | 440 | 361 | 152 | 240 | 161 | 199 | 98 | 108 | 103 | 132 | 45 | 37 | 19 | 13 |
| % HR diff | 65.8% | 68.6% | 71.5% | 67.1% | 60.8% | 64.0% | 78.4% | 66.3% | 62.0% | 69.9% | 52.3% | 62.2% | 73.0% | 89.5% | 69.2% |

each of the top-five rank positions (using the query-URL feature aggregation strategy, which had the best performance in the previous section), across each top-level ODP category.

Topics with little effect on NDCG from HR signals, or strong negative changes, are especially interesting since they deviate from the overall trends in Table 3. The "Reference" and "Society" topics may not stimulate a strong physiological response. Searches related to "Health" and "News" may be associated with a heightened HR before the click (e.g., in the case of health anxiety or reaction to a news headline), which directly contradicts the HR profile shown in Figure 1 (where the HR is expected to increase only after the result click has occurred). The extent of the per-topic differences suggest that we need further studies on how search topic affects HR.

## 4 DISCUSSION AND IMPLICATIONS

We have demonstrated the potential of large-scale physiological sensing to signficantly improve retrieval performance. The strong performance was obtained using longitudinal data from a relatively small set of 23k users for whom HR data was consistently available. The learned ranking model was not personalized, offering a considerable advantage: it could scale to all search engine users.

Looking ahead, we will focus on personalized result ranking [3] for a small set of Bing users for whom rich physiological data are available. We focused on heart rate given its broad availability across many wearables. Future work should explore more physiological signals, e.g., galvanic skin response, skin temperature, and the utilization of signals from different wearables.

We found topic-dependent performance differences, meaning that selective application of the ranking model could be valuable. Although the findings rely on being able to track physiology from many users, the magnitude of the observed gains may make it viable for search engines to recruit panels of wearable device users to collect large quantities of physiological and search data. Targeted recruitment of individuals with specific interests, HR profiles, or usage patterns, and learning dedicated per-cohort ranking models, may help further amplify the relevance gains we observed.

## REFERENCES


[1] Ioannis Arapakis, Konstantinos Athanasakos, and Joemon M Jose. 2010. A comparison of general vs personalised affective models for the prediction of topical relevance. In *SIGIR*. 371–378.

[2] Paul N Bennett, Krysta Svore, and Susan T Dumais. 2010. Classification-enhanced ranking. In *WWW*. 111–120.

[3] Paul N Bennett, Ryen W White, Wei Chu, Susan T Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the impact of short-and long-term behavior on search personalization. In *SIGIR*. 185–194.

[4] Paul Ekman, Robert W Levenson, and Wallace V Friesen. 1983. Autonomic nervous system activity distinguishes among emotions. *Science* 221, 4616 (1983), 1208–1210.

[5] Manuel JA Eugster, Tuukka Ruotsalo, Michiel M Spapé, Ilkka Kosunen, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. 2014. Predicting term-relevance from brain signals. In *SIGIR*. 425–434.

[6] Henry A Feild, James Allan, and Rosie Jones. 2010. Predicting searcher frustration. In *SIGIR*. 34–41.

[7] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems* 23, 2 (2005), 147–168.

[8] Jacek Gwizdka. 2014. Characterizing relevance with eye-tracking measures. In *IIiX*. 58–67.

[9] Irit Hochberg, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Elad Yom-Tov. 2016. A reinforcement learning system to encourage physical activity in diabetes patients. *arXiv preprint arXiv:1605.04070* (2016).

[10] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446.

[11] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *SIGKDD*. 133–142.

[12] Diane Kelly and Nicholas J Belkin. 2004. Display time as implicit feedback: understanding task effects. In *SIGIR*. 377–384.

[13] Liadh Kelly and Gareth JF Jones. 2010. Biometric response as a source of query independent scoring in lifelog retrieval. In *ECIR*. 520–531.

[14] Oscar D Lara and Miguel A Labrador. 2013. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials* 15, 3 (2013), 1192–1209.

[15] Yashar Moshfeghi and Joemon M Jose. 2013. An effective implicit relevance feedback technique using affective, physiological and behavioural features. In *SIGIR*. 133–142.

[16] Yashar Moshfeghi, Peter Triantafillou, and Frank E. Pollick. 2016. Understanding Information Need: An fMRI Study. In *SIGIR*. 335–344.

[17] Ryen W White and Steven M Drucker. 2007. Investigating behavioral variability in web search. In *WWW*. 21–30.

[18] Qiang Wu, Christopher JC Burges, Krysta M Svore, and Jianfeng Gao. 2010. Adapting boosting for information retrieval measures. *Information Retrieval* 13, 3 (2010), 254–270.