Engagement Periodicity in Search Engine Usage: Analysis and its Application to Search Quality Evaluation

Alexey Drutsa
Yandex
Moscow, Russia
adrutsa@yandex.ru

Gleb Gusev
Yandex
Moscow, Russia
gleb57@yandex-team.ru

Pavel Serdyukov
Yandex
Moscow, Russia
pavser@yandex-team.ru

ABSTRACT

Nowadays, billions of people use the Web in connection with their daily needs. A significant part of the needs are constituted by search tasks that are usually addressed by search engines. Thus, daily search needs result in regular user engagement with a search engine. User engagement with web sites and services was studied in various aspects, but there appear to be no studies of its regularity and periodicity. In this paper, we studied periodicity of the user engagement with a popular search engine through applying spectrum analysis to temporal sequences of different engagement metrics. We found periodicity patterns of user engagement and revealed classes of users whose periodicity patterns do not change over a long period of time. In addition, we used the spectrum series as metrics to evaluate search quality.

Categories and Subject Descriptions:
H.1.2 [User/Machine Systems]: Human information processing; H.5.2 [User interface]: Evaluation/methodology.

General Terms: Measurement, Experimentation

Keywords: User engagement; periodicity; DFT; spectrum analysis; quality metrics

1. INTRODUCTION

Some aims of users navigating the Web arise extremely rarely, some others arise repeatedly and make users access certain web services on a regular basis. The permanent usage of a particular service usually refers to a *user engagement* with it. The engagement is one of the major aspects [18] of user behavior that affects the profit and the popularity for majority of web sites. Thus, it is highly important for a web service to know and understand how its users are engaged. In order to describe and quantify user engagement, a series of metrics are developed. They could reflect different aspects of user engagement, from loyalty (e.g., the number of sessions per user per day [21]) to activity (e.g., the number of visited web pages [15]).

Web service usage is closely related to the individual life cycles of users (work, entertainment, holidays, etc.) that should result in periodicity of user engagement metrics. However, to the best of our knowledge, the effects of repetition and periodicity in user engagement are largely understudied in the existing research on user engagement (see Section 2).

In the current paper, we focus on periodicity in search behavior of users engaged by a search engine. On the one hand, one week and one year periodicity of behavioral metrics averaged over all users is well known[1]. On the other hand, the periodicity patterns for each individual user could vary significantly and may not match with the average periodicity [14]. For instance, an office worker uses a search engine for her work-related purposes at her work place. Therefore, her engagement with the search engine will be observed only within working days with different intensity which depends on the worker's schedule. On the other hand, a student may intensify her search engine usage during preparations for a test that could be held once every two or more weeks. Thus her search periodicity pattern differs from the pattern of the office worker. This motivates us to develop a methodology that detects periodicity phenomena in user engagement, and, in our work, we study and apply a technique that extracts individual periodic patterns of search engine users.

In order to catch periodicity in user engagement variance, we apply *spectrum analysis*[2]. Namely, we regard basic engagement metrics of a user (obtained from internal logs of a popular search engine) as time series within 4-week periods. Then, we translate them into sequences of Fourier amplitudes in a *frequency domain* by means of *the discrete Fourier transform* (DFT). We perform a detailed analysis of the obtained data and we show that there are identifiable patterns of user periodic behavior. Then we study the changes of these patterns over a period of five months. Further, we find stable groups of users that do not change their 4-week periodicity pattern during the long time period. We also show that the user engagement metrics extracted from the frequency domain are more consistent over the five-month period than the ones extracted from the source time domain. All these results justify the positive effect of applying the DFT for extracting periodicity from the user behavior data and *represent the first major contribution of this study.*

We reinforce the results of our analysis by applying the proposed method to evaluation of changes in a search sys-

WSDM'15, February 2–6, 2015, Shanghai, China.
Copyright 2015 ACM 978-1-4503-3317-7/15/02 ...$15.00.
http://dx.doi.org/10.1145/2684822.2685318.

---

[1] It is noted both in related work [14, 16, 20, 22, 24, 17] and observed in our study (see Section 5).

[2] Spectrum analysis is commonly used for identification of periodic phenomena in temporal sequences [23, 22, 24].

tem. Development of a search engine is based on ongoing updates, which are shipped permanently. In order to validate the positive effect of a change or at least to prevent its negative consequences, the development team compares the updated version of the service with respect to the previous one. The comparison is made in terms of quality, which needs to be quantified. This motivates to develop and study new search engine *quality metrics*. Variations in the ranking algorithm or the user interface (UI) may significantly change the habitual ways of interaction with the service and, therefore, may affect user periodicity behavior, which could serve as an indicator of these changes. In order to examine this assumption, we apply our periodicity patterns as new quality metrics (56 in total) and we validate them on a series of experiments that show (a) their insensitivity to changes, where updated version coincide with the previous one and (b) their sensitivity to artificial deterioration of the quality of different components of the search engine (such as the ranking algorithm, the UI, and the engine efficiency). In our experiments, we show that frequency domain approach for extracting periodicity patterns can be effectively used to provide more sensitive quality metrics for search engine evaluation. *We regard this as the second major contribution of our study.*

The rest of the paper is organized as follows. In Section 2, the related work is presented. In Section 3, the periodicity patterns of individual user behavior are introduced and the research questions are stated. In Section 4, we describe our user engagement data. In Section 5, the basic analysis of the periodicity patterns is presented. In Section 6, the long-term changes of user periodicity behavior are studied. In Section 7, the periodicity patterns are applied as metrics for search quality evaluation. In Section 8, the study's conclusions and future work are presented.

2. RELATED WORK

We compare our research with other studies in three aspects. The first one relates to search engine usage studies. The second one concerns periodicity in user behavior. The third aspect refers to analysis of user engagement with web services in general.

Search engine usage studies. On the one hand, the search engine usage studies on a high level could be divided with respect to short-term (within a search task or a search session) and long-term (across a series of such tasks and sessions) usage. On the other hand, they could be regarded both with respect to user interaction with the service and with respect to user's profit from usage of the service[3]. All these aspects are closely interrelated. For example, user dissatisfaction with a particular attempt to solve a search task may lead to a switch [25, 7, 4, 19] to another search engine on the short-term level. Simultaneously, a series of satisfied interactions with the service lead to the notion of user happiness [6, 9, 8, 10] on the long-term level. In our work, we study user engagement which is the long term user interaction component of the described classification.

Existing studies of user engagement with a search engine are three-fold. First, the studies [26, 11] discovered the relationship between search success and search engine reuse with respect to three search engines of different popularity. The authors of [26] also identified three groups of behavior patterns of simultaneous usage of the search engines within 6-month period (users who use only one search engine, who change search engines periodically, and who switch to another search engine and do not return back). In our work, we study 4-week behavior patterns, and their tendency to be stable during 5-month period is identified.

Second, some studies focused on the prediction of future changes in some user engagement metrics. Prediction of user switching type (no switch, persistent switch, or oscillating switch) between search engines during 26 weeks was studied in [26]. The authors of [21] developed a binary classifier to predict user engagement increase/decrease in the future week. They utilized the average number of sessions as the primary engagement metric on a par with some non-engagement measures (query types, user satisfaction, etc.) from the last three weeks.

Third, user engagement is used as service quality metrics for search engine evaluation. The search engine quality can be evaluated on different levels: the quality of a particular service component (e.g., a ranking algorithm, the user interface, etc.) and the quality of the web service as a whole either on short-term interactions (e.g., by evaluating search task satisfaction, switching, etc.) or on long-term usage (e.g., changes in happiness and engagement). The authors of [21] evaluated different changes in search relevance of a popular search engine by means of the A/B testing methodology with respect to the average number of sessions and several non-engagement measures reflecting query types and user satisfaction. The absence time (the time between two user visits) on a par with other engagement metrics was applied [5] to compare different ranking algorithms used at Yahoo! Answers by means of survival analysis. This methodology was applied in the recent study [2] to evaluate a web search engine changes in its ranking algorithm and its user interface. In our study, we utilize 56 new features resulted from the DFT of 4 engagement measures in order to evaluate, by means of A/B tests, different changes of the search engine ranking algorithm, changes of the user interface, and changes of the engine efficiency (see Section 7).

Periodicity in user behavior. To the best of our knowledge, no existing study investigated periodicity of individual user engagement. The existing works refered to the periodicity problem as seasonal factor (i.e., between years [20, 14]) or as weekday specific factor (i.e., between weeks [16, 17]). In both cases the term "periodicity" was mentioned either as a component of trend models (seasonal-trend decomposition methods [20, 16] or autocorrelation techniques [27, 17]) or as a variability of user behavior within a week (which was taken into account by introducing the day of the week as additional variable of models [5] and by providing different models for weekdays and weekends separately [16, 14]). In our work, we apply the DFT [23] to user engagement time series in order to encode the periodic behavior of an *individual user*[4]. Then, deep short-term and long-term analysis of user engagement periodicity are provided (see Sections 5 and 6 respectively).

User engagement with web services. The user engagement was studied not only for search engines, but also

---

[3] Actually, these classifications could be applied to any user need and to any product or service satisfying the need (not necessarily a search need).

[4] The DFT was applied to time series of query popularity (*aggregated over all users*) [22, 17, 24] and was used to detect dominant periods of the popularity variation over time.

for a wide range of web sites. The time between two successive user visits was used to understand differences in behavioral, content, and structural characteristics of web sites [1]. The authors of [16] compared different web sites with respect to user engagement metrics of popularity, activity, and loyalty that are aggregated over users for each web site. Using these metrics they obtained some engagement patterns (models) of the studied web sites. The research described in [15] is also devoted to comparison of a web site group with respect to multitasking user behavior, which analysis takes into account backpaging and tab switching during a browsing session. They used the number of page views and the dwell time (the presence time on a web site), both on visit and session levels, as the main engagement metrics of user activity. In our work, we study individual periodic behavior patterns of users (not web sites) based on both loyalty and activity aspects of their engagement.

3. FRAMEWORK AND NOTATIONS

In this section, we introduce methods of spectrum analysis [23] (also known as Fourier analysis) whose main component is the *discrete Fourier transform* (*DFT*). We remind the key points of the DFT in the next subsection. After that we motivate and discuss its application to the problem of extracting patterns of user engagement periodicity.

3.1 Discrete Fourier transform

Let $\mathbf{x} = (x_0, x_1, .., x_{N-1})$ be a sequence of N complex numbers, i.e., a vector in the space $\mathbb{C}^N$. Let $\{\mathbf{f}^k\}_{k=0}^{N-1}$ be the basis, where each basis vector $\mathbf{f}^k$ is the harmonic (sine wave) with coordinates $f_n^k = e^{\mathbf{i}\omega_k n}/N$, $n \in \mathbb{Z}_N$[5]. The values $\omega_k = 2\pi k/N$, $k \in \mathbb{Z}_N$, are called the *Fourier frequencies* (angular frequencies). Then, the *discrete Fourier transform* (*DFT*) of the sequence $\mathbf{x}$ is the sequence of its coordinates in the harmonic basis $\{\mathbf{f}^k\}_{k=0}^{N-1}$, that is, the series of complex numbers

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\mathbf{i}\omega_k n}, \quad \omega_k = \frac{2\pi k}{N}, \quad k \in \mathbb{Z}_N. \tag{1}$$

So, the DFT is a bijective map from *the source domain* into *the frequency domain* $\mathcal{F}(\mathbf{x}) = (X_0, .., X_{N-1})$, $\mathcal{F} : \mathbb{C}^N \leftrightarrow \mathbb{C}^N$.

Each complex number X_k could be represented in the form $X_k = |X_k| e^{\mathbf{i}\varphi_k}$, $k \in \mathbb{Z}_N$, where $|X_k|$ is the absolute value of the complex number and φ_k is its argument. The sequence of the absolute values normalized by N

$$\mathcal{A}(\mathbf{x}) = \Big(\frac{|X_0|}{N}, .., \frac{|X_{N-1}|}{N}\Big), \quad \mathcal{A} : \mathbb{C}^N \to \mathbb{R}_+^N, \tag{2}$$

is called *the amplitude series* (the amplitude vector or periodogram w.r.t. ω_k [23]). The sequence of the arguments $\Phi(\mathbf{x}) = (\varphi_0, .., \varphi_{N-1})$, $\Phi : \mathbb{C}^N \to [0, 2\pi)^N$, is called *the phase vector*. The amplitude $\mathcal{A}(\mathbf{x})_k$ represents the magnitude of the sine wave with the frequency ω_k presented in the series $\mathbf{x}$, whereas the phase $\Phi(\mathbf{x})_k$ represents how this wave is shifted.

If $\mathbf{x} \in \mathbb{R}^N$, then $X_0 \in \mathbb{R}$, and the DFT of $\mathbf{x} \in \mathbb{R}^N$ possesses the symmetry property: $X_k = \overline{X}_{N-k}$, $k = 1, .., [N/2]$[6]. If $\mathbf{x} \in {\mathbb{R}_+}^N$ (the case of our study), then $X_0 \geq 0$ and $|X_k| \leq X_0$, $k \in \mathbb{Z}_N$. Thus, without loss of generality, from here on in the paper the investigation of the amplitude series $\mathcal{A}(\mathbf{x})$ is replaced by investigation of its truncated part $\mathcal{A}(\mathbf{x}) = (\mathcal{A}(\mathbf{x})_0, .., \mathcal{A}(\mathbf{x})_{[N/2]})$ (using the same notation).

3.2 Periodicity patterns

Further we discuss some useful properties of the studied transformation. Let us consider some examples. Suppose that the time series $(0, .., 0, x_n, 0, .., 0)$, $n \in \mathbb{Z}_N$, has only one non-zero element (e.g., it represents a single user activity), then its amplitude vector is constant $(|x_n|/N, ..., |x_n|/N)$ independently of the position n of the non-zero value x_n in the series (e.g., see Fig. 2, col. "cS6"). In other words, the time information (position) of the single activity is accommodated in the phase component, whereas the activity type is bear by the amplitude component. The second example is provided by a time series, which is nearly constant (e.g., it represents a permanent constant user activity), i.e., $\mathbf{x} = (a + \varepsilon_0, .., a + \varepsilon_{N-1})$, $|\varepsilon_n| \ll |a|$, $n \in \mathbb{Z}_N$. In this case, the amplitude vector is of the form $(|a| + \delta_0, \delta_1, ..., \delta_{N-1})$ with small components $|\delta_k| \ll |a|$, $k \in \mathbb{Z}_N$ (e.g., see Fig. 2, col. "A1"). So, the amplitude vector carries the proportions between the magnitudes of the sine waves of different periodicity and disregards their shifts encoded by phases. Thus, the amplitude vector is the main component of the DFT for our study of user periodicity[7].

The main goal of our study is to catch patterns of user engagement periodicity. In order to separate this periodicity from the total amount of user engagement, we normalize the amplitude series by its zeroth component and obtain the normalized amplitude series $\mathcal{A}_{\mathcal{N}}(\mathbf{x}) = \mathcal{A}(\mathbf{x})/\mathcal{A}(\mathbf{x})_0$, which we refer to as *the periodicity pattern*[8]. Note that the normalized amplitudes are equal to the amplitudes of the normalized source series, i.e., the equality $\mathcal{A}_{\mathcal{N}}(\mathbf{x}) = \mathcal{A}(\mathbf{x}/\|\mathbf{x}\|_1)$ holds, where $\|\mathbf{x}\|_1 = \sum_{n=0}^{N-1} |x_n|/N$ is the l_1-norm of $\mathbf{x}$. Thus, we have $\mathcal{A}_{\mathcal{N}}(\mathbf{x}) = \mathcal{A}_{\mathcal{N}}(\mathbf{y})$, if $\mathbf{x} = \lambda\mathbf{y}$. It means that the normalized amplitude $\mathcal{A}_{\mathcal{N}}(\mathbf{x})$ captures periodicity of $\mathbf{x}$ with respect to its scale and does not distinguish between two source series that differ by a multiplicative constant only.

Series under study. In our work, we study time series of different user engagement features and their transformations. Given a user's engagement feature, under our study are the following series: (a) $\mathbf{x}$ is *the source time series* of the feature of length N; (b) $\mathcal{A}(\mathbf{x})$ is *its amplitude series* of length $[N/2]+1$; and (c) $\mathcal{A}_{\mathcal{N}}(\mathbf{x})$ is *its periodicity pattern* (or normalized amplitude series) of length $[N/2]$.

3.3 Research questions

The main goal of our study is to identify the benefit of spectrum analysis methodology for analysis of user engagement (and particularly its periodicity). Thereupon, we translate this objective into the following research questions:

- **[RQ1]** Could clusters of users with common periodicity behavior be identified?
- **[RQ2]** Could users with stable periodicity be identified?
- **[RQ3]** Could the features of spectrum analysis be used with profit in a practical problem faced by a search engine?

---

[5] We remind that $e^{\mathbf{i}\varphi} = \cos\varphi + \mathbf{i}\sin\varphi$, $\mathbf{i}$ is imaginary unit, and $\mathbb{Z}_N = \{0, 1, .., N-1\}$.

[6] $\overline{z} = a - \mathbf{i}b$ is the complex conjugate of $z = a + \mathbf{i}b$.

[7] A study of phases is left for future work.

[8] Since we always have $\mathcal{A}_{\mathcal{N}}(\mathbf{x})_0 = 1$, we assume that the series $\mathcal{A}_{\mathcal{N}}(\mathbf{x})$ begins with $\mathcal{A}_{\mathcal{N}}(\mathbf{x})_1$ from here on in the paper.

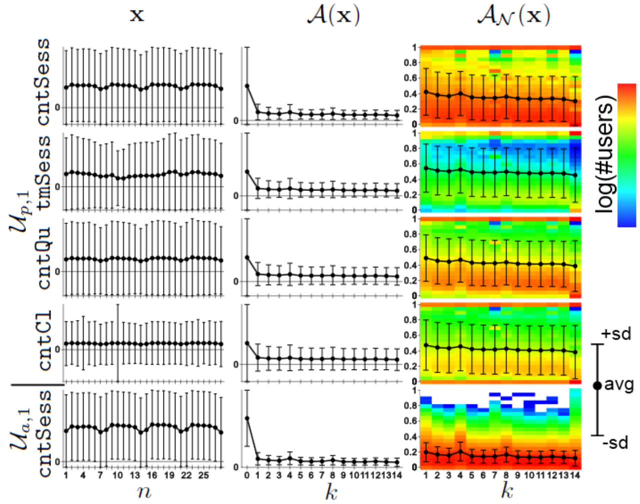

Figure 1: The distribution (*avg*, *sd*, and the heat map) of users from the sets $\mathcal{U}_{p,1}$ and $\mathcal{U}_{a,1}$ w.r.t. each component of the source time series x, the amplitude series $\mathcal{A}(\mathbf{x})$, and the periodicity pattern $\mathcal{A}_{\mathcal{N}}(\mathbf{x})$ for each engagement measure.

4. THE DATA

Engagement measures. For each user, we study 4 basic additive measures, which represent both loyalty and activity aspects of user engagement. They are (aggregated over a day):

- the number of sessions (`cntSess`);
- the presence time (sum of session times, `tmSess`);
- the number of queries (`cntQu`);
- the number of clicks (`cntCl`).

The measure `cntSess` corresponds to the loyalty aspect of user engagement, whereas the measures `tmSess`, `cntQu` and `cntClick` correspond to user activity [5, 18, 21].

Time periods. We take the logs of Yandex[9], one of the most popular search engines, from the 1st September, 2013 to 15th March, 2014 (196 days). For each user, we split her actions (i.e., interactions with all web services of Yandex) to sessions[10]. Using this data, for each day in the time period and for each user, we calculate number of sessions, presence time, number of queries, and number of clicks that she produced in that day. Then we define 28-day period (4 weeks) as the basic time length of the source sequences under study (i.e., $N = 28$)[11]. We split the data sequences to 7 consecutive 28-day periods $\{\mathbb{T}_i\}_{i=0}^{6}$, $|\mathbb{T}_i| = 28$.

Sets of users[12]. Then, for each $i = 1, .., 5$, we build the set $\mathcal{U}_{p,i}$ consisting of users that have at least one action during each of the time periods $\mathbb{T}_{i-1}$, $\mathbb{T}_i$, and $\mathbb{T}_{i+1}$[13]. For a deeper study on periodicity, we also consider the *set of active users* $\mathcal{U}_{a,i} \subset \mathcal{U}_{p,i}$ $(i = 1, .., 5)$ formed by those users that have actions at 14 days of the period $\mathbb{T}_i$ at least. The user set $\mathcal{U}_{a,i}$ consists of around 33-39% of the users from the set $\mathcal{U}_{p,i}$, but it makes 72-79% of their sessions, 74-83% of their presence time, 73-81% of their queries, and 74-82% of their clicks (depending on i). We use the intersection of the user sets $\mathcal{U}_p = \bigcap_{i=1}^{5} \mathcal{U}_{p,i}$ (a user is presented in each time period) and $\mathcal{U}_a = \bigcap_{i=1}^{5} \mathcal{U}_{a,i}$ (a user is active in all time periods) for long-term studies. Each of the studied user sets contains several millions of users.

Thus, for each engagement measure (`cntSess`, `tmSess`, `cntQu`, and `cntCl`), for each time period $\mathbb{T}_i, i = 1, .., 5$, and for each user belonging to the set $\mathcal{U}_{p,i}$, we have the 28-length source time series of the engagement measure. So, each user may possess up to 20 source time series, each of them is processed through the DFT providing the 15-length amplitude series and 14-length periodicity patterns.

5. SPECTRUM ANALYSIS

In this section, we provide the analysis of the periodicity patterns within one time period. First, we provide basic analysis of the studied series for our data sets. Second, we identify groups of users that represent similar periodicity behavior. The results are presented for the time period $\mathbb{T}_1$, the ones for the other periods are analogous. The relationships among the periods are studied in the next section.

5.1 Basic distributions of the series

In Fig. 1, we present[14] the distribution (the average value *avg*, the standard deviation *sd*, and the heat map) of users from the set $\mathcal{U}_{p,1}$ with respect to each component of the source time series $\mathbf{x}$, the amplitude series $\mathcal{A}(\mathbf{x})$, and the periodicity pattern $\mathcal{A}_{\mathcal{N}}(\mathbf{x})$ for each of the 4 engagement measures. We see that the components of the source time series $\mathbf{x}$ vary considerably among users. The loyalty user engagement measure is less oscillating than the activity ones: the average values of normalized amplitudes $\mathcal{A}_{\mathcal{N}}(\mathbf{x})$ are greater than 0.4 for (`tmSess`, `cntQu`, and `cntCl`) and are lower than 0.4 for `cntSess`. However, the distributions' shapes do not differ noticeably among engagement measures. Therefore, due to the space constraints, several results discussed further are reported only for one engagement measure (usually for `cntSess`), when they are similar for the others.

The same results are presented at the bottom of Fig. 1 for the active users $\mathcal{U}_{a,1}$ (they interact with the search engine at least 14 days in the time period) only for the number of sessions (`cntSess`). The main difference from the set $\mathcal{U}_{p,1}$ is seen at the top of its heat map: the periodicity patterns are noticeably lower 1. Hence, some other observations made for the set $\mathcal{U}_{p,1}$ are more clearly visible for the active users: the distributions of the normalized amplitudes $\mathcal{A}_{\mathcal{N}}(\mathbf{x})$ are shifted closer to 0 and smoother (see also Fig. 3), the peaks at frequencies $k \in \{4, 8\}$ are more dominating among others.

The peaks in the distributions of the $\mathcal{A}_{\mathcal{N}}(\mathbf{x})_k$ for the set $\mathcal{U}_{p,1}$ (see Fig. 3(p)) correspond to the clusters of periodicity patterns (further in our study) whose centroids are presented in Fig. 4. As well, we observe analogous steps on the dis-

---

[9] yandex.com

[10] A session is commonly defined as a sequence of actions whose dwell times are less than 30 minutes [2, 5, 12, 21].

[11] This length allows to catch periodicity between weeks and, meanwhile, study its long-term changes as we further describe.

[12] In this paper, we use cookie IDs to identify users as done in other studies on user engagement [5, 15, 21].

[13] Thus, we eliminate the impact on our periodicity patterns from the effects, when a user can start for the first time or finish at all using the search engine in the studied time period $\mathbb{T}_i$ (e.g., a user clears its browser cookies). It is about 24-27% of all users depending on i.

[14] From here on in the paper we hide all absolute values for confidentiality reasons.

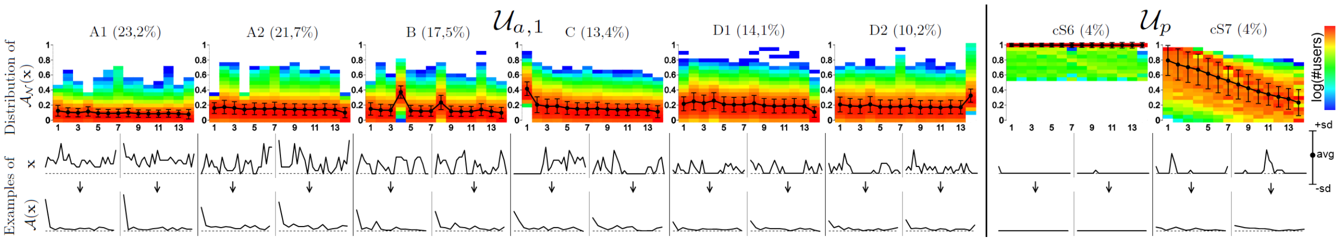

Figure 2: The distribution and examples of users from several periodicity models (A1, A2, B, C, D1, D2, cS6, and cS7) w.r.t. the number of sessions `cntSess` (see Section 5.2 for details).

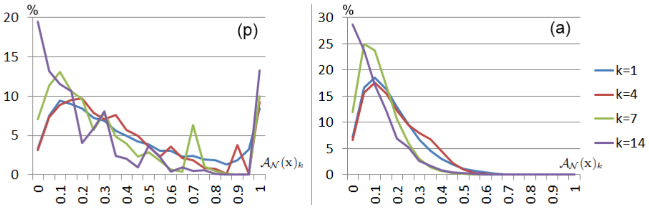

Figure 3: Density of distribution of users $\mathcal{U}_{p,1}$ (p) and $\mathcal{U}_{a,1}$ (a) w.r.t. the value $\mathcal{A}_\mathcal{N}(\texttt{cntSess})_k, k = 1, 4, 7, 14$.

tributions of the active users $\mathcal{U}_{a,1}$ (e.g., see Fig. 3(a) for the freq. $k = 4$ near the value 0.36 and $k = 14$ near 0.25), they are also detected further in our study (see Fig. 2).

The weekly periodicity is clearly observed on the values of the source time series averaged over users. This periodicity is also detected by the periodicity patterns and is expressed by the peaks on the amplitude series at frequencies $k = 4$ and 8[15]. These frequencies are responsible for the weekly periodicity, as it is seen for users whose activity drops at weekends (see Fig. 2, col. "B"). A single day interaction type with the web service (as examples in Fig. 2, col. "cS6") could not be detected by the average values of the series, while the heat maps of periodicity patterns detect such type of activity (Fig. 1, $\mathcal{A}_\mathcal{N}(\mathbf{x})$ are near 1) and show a considerable amount of such users (see also several slices of distribution in Fig. 3(p)). Thus, our periodicity patterns are able to catch and encode basic periodicity types. Further in our paper, we show how the periodicity patterns detect more complicated activity types, which are difficult to find in the source time series with the naked eye.

5.2 User engagement periodicity models

In this subsection, we identify groups of users which have similar periodicity behavior. They are referred to as *the periodicity models* of user engagement. We utilize standard k-means clustering algorithm with respect to the periodicity patterns $\mathcal{A}_\mathcal{N}(\mathbf{x})$ that were treated as vectors of 14-dimensional space of normalized amplitudes. We run it 10 times with random seed means for different numbers of clusters K (from 2 to 30). Further we report the most interesting results only. We find that the most sensible and stable clusters are detected for $K = 7$ (for all users) and for $K = 6$ (for active users)[16].

We present the results of the clustering for the active users $\mathcal{U}_{a,1}$ in Fig. 2 with respect to the number of sessions `cntSess`

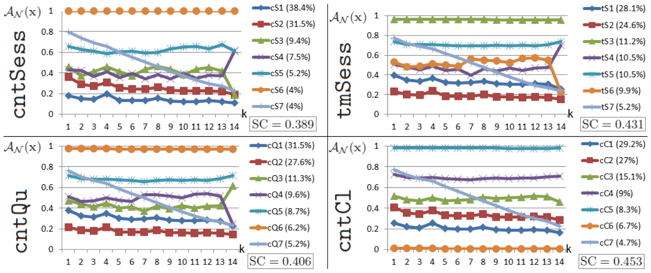

Figure 4: The 7 cluster centroids for each engagement measure across users $\mathcal{U}_p$.

(the silhouette coefficient[17] SC = 0.322): there are the periodicity models **A1** (23.2%), **A2** (21.7%), **B** (17.5%), **C** (13.4%), **D1** (14.1%), and **D2** (10.2%). Each group is supplied by the heat map of its periodicity pattern distribution and two example users with its source time and amplitude series which are close to the centroids. The periodicity models **A1**, **A2**, **B**, and **C** have strong and coherent interpretations. The users from **A1** (*the permanent model*) interact with the search engine permanently and, thus, their periodicity patterns are small (i.e., all sine waves are smaller than a constant). The users of **A2** behave similarly (together with **A1** they form 44.9% of all users), but their variation in interaction activity is higher. The users from **B** (*the office worker model*) interact with the web service with week-periodicity, which strongly reflects in the domination of the 4-th and 8-th components of their periodicity patterns (i.e., the sine waves with the periods multiple of 7 days dominate among others). The users from **C** (*the holiday model*) have up to 10-day drop in search engine usage[18], which is reflected in the domination of the 1-st component of the periodicity patterns (i.e., the sine wave with the period of 28 days dominates others). The activities of users from **D1** and **D2** have both large variation and have no sensible periodic interpretation. The relationship between all these periodicity models during the 20-week time period is studied in Section 6.

We present the centroids and the silhouette coefficients (SC) of 7-clustering of all users $\mathcal{U}_p$ (not only active, as before) with respect to periodicity patterns of different user engagement measures in Fig. 4. First, we see that the obtained periodicity models are of similar form in vast majority of the cases. The only one exception is the group **cC6**, which represents the users that have no clicks at all (i.e., the source time series of the number of clicks consists of zero values). The other user engagement measures cannot have zero time-

---

[15] From the DFT properties, it follows that the amplitudes $\mathcal{A}(\mathbf{x})_k = 0$ for k not divisible by 4, if the series $\mathbf{x}$ is periodical with period 7 ($N = 4 \cdot 7$).

[16] For these K, all 10 runs converges to the same centroids. Greater K results in different centroids for different seed means.

[17] en.wikipedia.org/wiki/Silhouette_(clustering)

[18] We remind that the absence of a user activity at the start or at the end is not caused by their start or end of usage of the search engine (see Section 4)

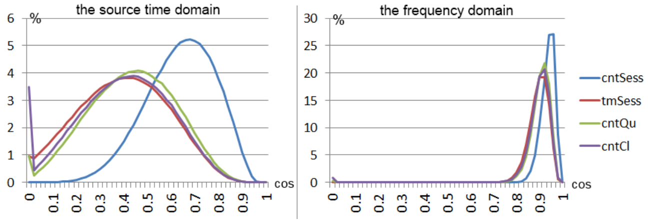

Figure 5: Density of distribution of users $\mathcal{U}_a$ w.r.t. theirs correlation coefficients between the time periods $\mathbb{T}_1$ and $\mathbb{T}_2$ for the source time series and for the amplitude series.

series, thus the analogous cluster is not detected for them. Second, comparing the models of all users (w.r.t. the number of sessions `cntSess`) and the ones of the active users $\mathcal{U}_{a,1}$ (Fig. 2), we see that the latter ones do not include the clusters **cS5**, **cS6**, and **cS7**, due to the latter clusters represent the users with very low activity model. For instance, the periodicity models **cS6** and **cS7** are presented in Fig. 2, where the examples demonstrate which periodicity behavior are encoded by each model.

The basic and cluster analyses of user periodicity behavior, described in this section, allow us to conclude, that *the periodicity patterns provide strong identification of several large user clusters with common coherent periodicity behavior*. It is the answer to the **RQ1**.

6. LONG-TERM PERIODICITY CHANGES

In the previous section, we provided a deep analysis of the source time series and their periodicity patterns of individual users within a 4-week period. Now we study how the periodicity patterns change across a sequence of such periods, namely within a 20-week period.

6.1 Correlation between the series

First of all, we investigate how the user series (both the source time series and the periodicity patterns) are similar across periods $\{\mathbb{T}_i\}_{i=1}^{5}$ for the same user. We measure the series similarity in terms of correlation of series, i.e., the cosine of the angle between the series: $\cos(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{y})/(\|\mathbf{x}\|\|\mathbf{y}\|)$ for any non-zero series $\mathbf{x}, \mathbf{y}$, $(\mathbf{x}, \mathbf{y}) = \sum_n x_n y_n$ and $\|\mathbf{x}\|^2 = (\mathbf{x}, \mathbf{x})$. Thus, given a user, the engagement measure, the series type (the source time series or the periodicity pattern), and time periods $\mathbb{T}_i$ and $\mathbb{T}_j$, one has a correlation coefficient whose distribution over a set of users is investigated.

The distributions of users $\mathcal{U}_a$ w.r.t. their correlation coefficients between the time periods $\mathbb{T}_1$ and $\mathbb{T}_2$ for the source time series and for the amplitude series are presented in Fig. 5. The average values of the correlation coefficients for the source time series are equal to $0.65_{\text{sd}0.15}$ for `cntSess`, $0.41_{\text{sd}0.19}$ for `tmSess`, $0.44_{\text{sd}0.18}$ for `cntQu`, and $0.41_{\text{sd}0.19}$ for `cntCl`, whereas the ones for the amplitude series are $0.95_{\text{sd}0.03}$ for `cntSess`, $0.91_{\text{sd}0.04}$ for `tmSess`, $0.91_{\text{sd}0.07}$ for `cntQu`, and $0.90_{\text{sd}0.1}$ for `cntCl`. The results for all users $\mathcal{U}_p$ are less dramatic, but analogous. For instance, for the number of sessions `cntSess`, the average values are $0.39_{\text{sd}0.26}$ (the time domain) and $0.89_{\text{sd}0.06}$ (the frequency domain). Meanwhile, the distributions for the same series types (both in the time domain and the frequency domain) do not vary noticeably over the distance between the time periods. For instance, the average values among active users $\mathcal{U}_a$ for `cntSess` in the frequency domain are $0.945_{\text{sd}0.03}$ for the pair $\mathbb{T}_1 - \mathbb{T}_3$, $0.939_{\text{sd}0.03}$ for $\mathbb{T}_1 - \mathbb{T}_4$, $0.943_{\text{sd}0.03}$ for $\mathbb{T}_1 - \mathbb{T}_5$.

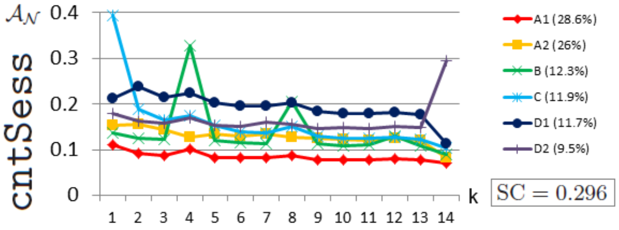

Figure 6: The 6 cluster centroids for active users $\mathcal{U}_a$.

These results show that the same time series data have significantly better correlation across months in the frequency domains (processed through the DFT) than in the time domain (in the source form) both for all users and for its active subset. In other words, the periodicity patterns catch those individual user behavior traits that do not change during a long time period.

6.2 Transitions between periodicity models

In order to investigate, how user's periodicity models change over a long time period, we apply the clustering algorithm (as in the previous section) to the active users $\mathcal{U}_a$[19]. The resulting cluster centroids are presented in Fig. 6. We see that the obtained periodicity models are of the same form as for $\mathcal{U}_{a,1}$ in Fig. 2, but differ in their sizes due to differences in the time periods (see further).

The transition map of users $\mathcal{U}_a$ between the six periodicity models (user groups) during 20-week time period ($\mathbb{T}_i, i = 1, .., 5$) is represented in Fig. 7. The nodes' radii are proportional to the sizes of the groups at each time period $\mathbb{T}_i, i = 1, .., 5$. The thickness of the lines between the nodes are proportional to the amount of users that switch between the models. Each transition $\mathbb{T}_i \to \mathbb{T}_{i+1}, i = 1, .., 4$, is supplied by a diagram of switching between the models (the bar graphs represent the percent of users that stay in the same cluster or switch to another one w.r.t. the number of all users). From Fig. 7, one could learn that there are 3 stable periodicity models (**A1**, **A2**, and **B**) across all periods except $\mathbb{T}_4$. At the time period $\mathbb{T}_4$, *the holiday model* **C** grows by significantly accumulating users from other groups and especially from the clusters **A1**, **A2**, and **B**. And further, at the time period $\mathbb{T}_5$, the usual distribution of users restores.

What happens at the time period $\mathbb{T}_4$? One could see that the periodicity model **C** corresponds to the periodicity pattern, where the frequency ω_1 dominates others (see Fig. 2, col. "C"). It corresponds to the users whose activity is absent at a continuous half of the time period (see Section 5.2). The time period $\mathbb{T}_4$ contains the Christmas and New Year holidays. So, we suppose that the transitions $\mathbb{T}_3 \to \mathbb{T}_4$ and $\mathbb{T}_4 \to \mathbb{T}_5$ catch users vacancies (holidays). While the groups **A1** (*the permanent model*) and **A2** lose about 12% and 17%, respectively, of their users at $\mathbb{T}_4$ for benefit of **C** (usually, they lose 7% and 11%, resp.), the outflow of users from *the office worker model* **B** to the cluster **C** raises up to 32%, being usually at 8.5% at different periods $\mathbb{T}_i$. The periodic structure of **B** corresponds to the weekly periodical activity (e.g., as of an office worker, see Fig. 2, col. "B"), that is, the users whose activity is weekly periodical and occurs mostly

---

[19] Each user from $\mathcal{U}_a$ possesses 5 periodicity patterns (one per each 4-week time period $\mathbb{T}_i, i = 1, .., 5$), each of these patterns is utilized in the k-means algorithm.

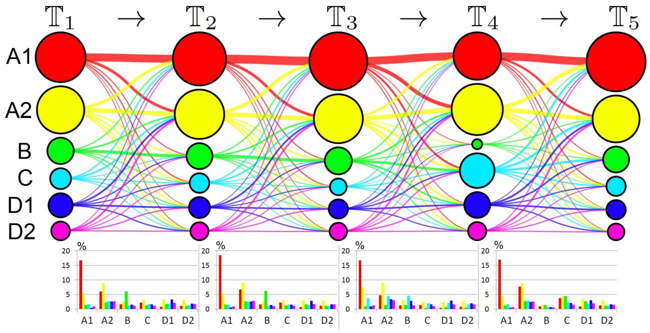

Figure 7: The transition map of the users $\mathcal{U}_a$ between the 6 periodicity clusters during the 20-week period ($\{\mathbb{T}_i\}_{i=1}^{5}$) and the diagrams for each transition $\mathbb{T}_i \to \mathbb{T}_{i+1}, i = 1, .., 4$ (see Section 6.2 for details).

within working days, are significantly subjected to vacation seasons. It is the second argument, which supports our supposition. So, people take a holiday during a year and the vast majority of them are taken near the Christmas, which is identified by the observed long-term changes of *the holiday model* **C**.

In order to identify users who do not change their periodicity models, we present in Table 1 col. **(I)** the proportions of users that have the same periodicity model across all 4-week time periods w.r.t. the size of the periodicity model group at the first 4-week period $\mathbb{T}_1$. We see that only *the permanent model* **A1** of the search engine usage has a significant fraction (28.6%) of users with stable periodicity model, the second one (by a large margin) is *the office worker model* **B**. Further, we join the periodicity models into sets of models and, for each set, we calculate the proportions of users whose periodicity model holds in the set across all 4-week time periods (i.e., users *oscillate between models* in the set) w.r.t. the sum of the sizes of the set's periodicity model groups at the first 4-week period $\mathbb{T}_1$. These proportions of users who oscillate between 2, 4, and 5 models across the 20-week period are presented in Table 1 at the bottom triangle under the diagonal of col. **(II)**, at the top triangle above the diagonal of col. **(II)**, and at col. **(III)** respectively. A cell in raw **X** and column **Y** of the upper triangle of col. **(II)** corresponds to the set of 4 models, which does not contain models **X**, **Y**. A cell in raw **X** from col. **(III)** corresponds to the set of 5 models, which does not contain model **X**. For instance, the cell **(B, D1)** reports that 56.4% of users who have periodicity models **A1**, **A2**, **C**, or **D2** (a set of models) at the first period $\mathbb{T}_1$ do not leave these 4 models in the next 4 periods. We see that the oscillation between **A1** and **A2** models is most stable among all 2-model oscillations. Meanwhile, we see that the best stable 5-model oscillation is the one which does not include *the office worker model* **B**, which represents weekly periodicity. Thus, we identified the following stable user behaviors: (a) *a user is likely to use search engine permanently if she has the same usage at the initial time period*; (b) *it is unlikely that user will behave as an office worker if she did not do it at the initial time period.*

Summarizing the results, first, we found that the periodicity behavior encoded in the periodicity patterns is significantly more stable in long-term perspective than the user behavior data presented in the time domain. Second, we found that the users with definitive models of periodicity behavior stick to one periodicity model across months and the changes in their models are mostly become caused by changes in their life style (e.g., taking Christmas holidays). Finally, we conclude that *the periodicity patterns catch individual user traits that are mostly stable during a long time period*. It is the answer to the **RQ2**.

Table 1: Fraction of users that have stable periodicity model across 20 weeks (I) and that oscillate between 2 (the bottom triangle of (II)), 4 (the top triangle of (II)), and 5 (III) models.

	(I)	(II)						(III)
		A1	A2	B	C	D1	D2	
A1	28.6%	—	26.1%	35.6%	27.7%	22.2%	30.1%	52.3%
A2	1.9%	36.6%	—	27.6%	24.5%	30.4%	32.8%	40.3%
B	3.3%	22.9%	5.5%	—	47.9%	56.4%	53.6%	74.4%
C	0.2%	27.7%	5.4%	11.8%	—	45%	45%	59.3%
D1	0.9%	20.2%	6.1%	7.9%	3.1%	—	55.2%	66.2%
D2	0.8%	22.7%	6%	3.4%	1.3%	4.5%	—	68.4%

7. SEARCH ENGINE QUALITY METRICS

In this section we show how the periodicity patterns of user engagement could be used as evaluation metrics of the search engine quality.

Quality evaluation problem. The quality could be measured at different stages of the search engine development processes and for changes in different components of the system. User engagement metrics are believed to be the most relevant to the company profit and they are able to quantify the quality of changes in any service component on the same measurement scale. Therefore, these metrics are the most universal and are used to qualify the appropriateness of lower-level quality metrics developed for more specific development tasks. A measure claimed to be used as a quality metric should be validated in a series of experiments. These experiments should show that (a) the deviation of the measure from zero is not significant in a comparison of two identical versions of the service and (b) the metric indicates a significant difference of two versions of the system (usually, the current production version and another one with some artificial deterioration).

In order to compare two states of the service, we apply the technique of *A/B experiments* (*A/B tests*) [3, 21, 13], which is widely used in search engine evaluation. The essence of the method consists in sampling of two groups of users (A and B) from the whole flow of users of the service. The users from the group A (*the control group*) and B (*the treatment group*) are exposed during some time period to the old service version and to the new one, respectively. After the experimentation time period, the quality metric is calculated for these users, and the relative difference of the average values [21] over the user groups is calculated:

$$\text{Diff} = \chi \cdot (\text{avg}_B(m) - \text{avg}_A(m))/\text{avg}_A(m)^{20}.$$

Nonetheless, the quantity Diff could not serve itself as an indicator of positive or negative consequences of the evaluated update of the search engine. The relative difference of the averaged values should be controlled by a statistical significance test. In our study, we apply two-sample t-test (as in [3, 21]) to decide weather the metric aggregated over users from the treatment group is significantly larger or smaller

---

[20] The factor χ is randomly chosen once in our study in order to hide real values for confidentiality reasons.

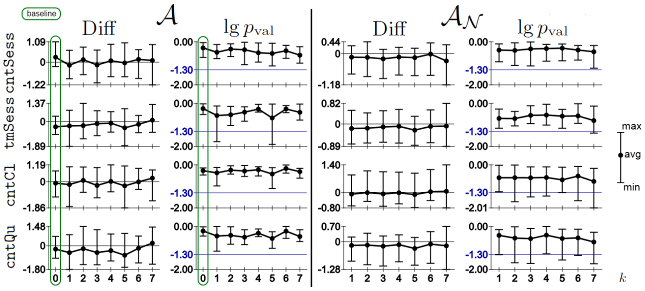

Figure 8: Results of 5 control experiments (A/A tests) for 4 user engagement features, for $\mathcal{A}$ and $\mathcal{A}_{\mathcal{N}}$ transformations (avg/max/min of Diff and $\lg p_{\mathrm{val}}$).

than that aggregated over users from control one. We considered a commonly used threshold $p_{\mathrm{val}} = 0.05$ for the p-value of the test. We also use decimal logarithm $\lg p_{\mathrm{val}}$ in our figures whose threshold is $\lg p_{\mathrm{val}} = \lg 0.05 \approx -1.3$.

Experiment setup. In our work, we conduct 25 A/B experiments in total: 3 for different changes in the ranking algorithm, 5 for the engine response time increase, 12 for different changes in the user interface, and 5 control experiments (so-called A/A tests [3]), where old and new versions of the system coincide. Each experimentation has 14-day time period, which is motivated by the following arguments. On the one hand, the experiment duration directly affects the speed of the decision to ship the service update, and, therefore, it should be as short as possible. On the other hand, the difference in behavior of control and treatment groups should be obtained on the desirable level of statistical significance, which also depends on the length of the period. The study [21] of user engagement metrics (`cntSess`) w.r.t. A/B tests shows that the metric aggregated over one day of experiment has poor significance level, and it is recommended [13] to accumulate metric values at 2-week period.

Our periodicity patterns and amplitudes of user engagement are believed to be a good extension of existing quality metrics based on user engagement, because they are alternative approaches of metric accumulation (besides the summation and averaging), and we rely on the following intuition. The periodicity pattern encodes variation of a user behavior over time. It tends to 0, if the variation is low, and tends to 1 otherwise (as it observed in Sections 5 and 6). Hence, a negative or positive change in a search engine component influences user behavior in *some of her search tasks*[21], thus, it changes user behavior variation and, therefore, her periodicity pattern. So, in this section (unlike the remainder of the paper), we use the DFT for the time series with the length $N = 14$ and, therefore, the truncated amplitude vector has the length $[N/2] + 1 = 8$. So, we investigate 56 *new scalar user engagement metrics*: for each user engagement feature $\{\texttt{cntSess}, \texttt{tmSess}, \texttt{cntQu}, \texttt{cntCl}\}$ (see Section 4) with the time series $\mathbf{x}$, we study the amplitudes $\mathcal{A}(\mathbf{x})_k$ and their normalized values $\mathcal{A}_{\mathcal{N}}(\mathbf{x})_k, k = 1, .., 7$ (the periodicity pattern). The amplitude $\mathcal{A}(\mathbf{x})_0$ is the average value of the time series $\mathbf{x}$, therefore, it serves as the baseline metric[22].

Experimental results. First of all, we present the results for 5 control experiments (A/A tests) in Fig. 8 (the average and the max/min values of Diff and $\lg p_{\mathrm{val}}$ w.r.t. 5 obtained values). We see that, for almost all metrics, the average value of Diff is near zero. Expectedly, the significance levels of A/A tests averaged over each metric are very low ($p_{\mathrm{val}} > 10^{-0.72}$). Meanwhile, some metrics catch a significant difference (the p-value < 0.05 or $\lg p_{\mathrm{val}} < -1.3$) between equal versions of the web service on one control experiment among 5: two amplitudes ($\{\mathcal{A}(\texttt{tmSess})_k\}_{k=1,5}$) and some normalized ones ($\mathcal{A}_{\mathcal{N}}(\texttt{tmSess})_7$, $\{\mathcal{A}_{\mathcal{N}}(\texttt{cntCl})_k\}_{k \neq 1,5}$, and $\{\mathcal{A}_{\mathcal{N}}(\texttt{cntQu})_k\}_{k \neq 4}$). The number of passed control experiments by each metrics is reported as the *first number* in the cells of Table 2.

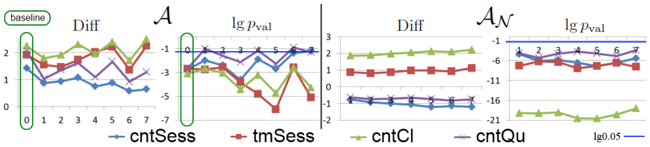

Figure 9: Results of the swap-2 ranking deterioration (A/B tests) for $\mathcal{A}$ and $\mathcal{A}_{\mathcal{N}}$ transformations.

Ranking quality evaluation. Now we apply our metrics to evaluate the search engine changes. We start from the ranking algorithm that is the main component of the service. In the first experiment, we swap two random results from the first quintuple with two random ones from the second quintuple in all results presented for the treatment group B (the "swap-2" experiment). The results of the experiment are presented in Fig. 9. One could see that all normalized amplitudes $\mathcal{A}_{\mathcal{N}}$ of all user engagement measures provide a very high significance level ($p_{\mathrm{val}} < 10^{-3.77}$) and strongly outperform the corresponding baseline average values $\mathcal{A}(\,\cdot\,)_0$ w.r.t. the significance level. Moreover, we learn that amplitudes $\mathcal{A}$ of the presence time `tmSess` and the number of clicks `cntCl` have also a high significance level ($p_{\mathrm{val}} < 10^{-2.56}$) and almost all of them outperform the baseline average values $\mathcal{A}(\,\cdot\,)_0$ by the significance level. The number of queries `cntQu` has some significant amplitudes, but all of them are outperformed by the average value. The same is valid for the number of sessions `cntSess`, but the amplitude $\mathcal{A}(\texttt{cntSess})_3$ outperforms the baseline metric.

We conduct two other experiments whose ranking deterioration are weaker than the one of the "swap-2" experiment. They are personalization switch off ("PersOff") and the swap of the second and the fourth results ("2-4-swap"). Unfortunately, only a few metrics catch significant differences in these experiments: they are the amplitudes $\mathcal{A}(\texttt{cntSess})_{k=3,7}$ and $\mathcal{A}(\texttt{cntQu})_7$ ($p_{\mathrm{val}} < 10^{-1.75}$), the normalized amplitudes $\mathcal{A}_{\mathcal{N}}(\texttt{cntSess})_{k=1,3}$, and $\mathcal{A}_{\mathcal{N}}(\texttt{cntQu})_1$ ($p_{\mathrm{val}} < 10^{-1.36}$) for "PersOff", and the amplitude $\mathcal{A}(\texttt{cntSess})_5$ ($p_{\mathrm{val}} < 10^{-1.303}$) for "2-4-swap". For the baseline average values $\mathcal{A}(\,\cdot\,)_0$, p-values do not pass the threshold of 0.05. Finally, we report the number of ranking experiments passed by each metric as the *first digit* of the *superscript* in the cells of Table 2.

Evaluation of UI changes. The second main component of any web service is the user interface (UI). In Fig. 10 and Fig. 11 we demonstrate the results for 7 A/B experiments on *heavy UI changes* and 5 A/B experiments on *light UI changes* correspondingly (the average and the max/min values of Diff and $\lg p_{\mathrm{val}}$). The heavy UI changes include deterioration of link colors, margins, opacity, and fonts of the search results, whereas the light changes include neutral variations of the UI. We see that a part of the heavy UI changes is caught by almost all metrics, however, amplitudes of the presence time `tmSess` and the number of clicks

---

[21] Usually a search engine update is noticeable in very small number of scenarios of interaction with the web service.

[22] The average value of `cntSess` was studied in [21].

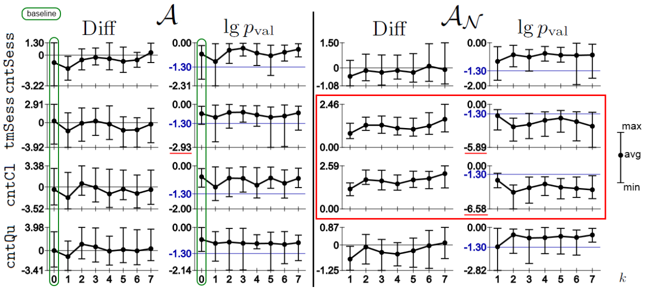

Figure 10: Results of 7 UI change experiments for 4 user engagement metrics, $\mathcal{A}$ and $\mathcal{A}_{\mathcal{N}}$ transformations (avg/max/min of Diff **and** $\lg p_{\text{val}}$**).**

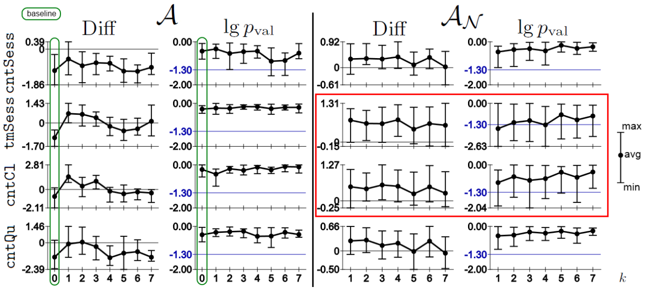

Figure 11: Results of 5 light UI change experiments for 4 user engagement features, $\mathcal{A}$ and $\mathcal{A}_{\mathcal{N}}$ transformations (avg/max/min values of Diff **and** $\lg p_{\text{val}}$**).**

`cntCl` have p-value higher than 0.1 at average. On the contrary, their normalized amplitudes catch all the changes in the UI on a highly significant level (outlined by red boxes in both figures) and outperform the baseline average metrics $\mathcal{A}(\cdot)_0$. As well, these 14 metrics detect light UI changes as significant, whereas the other metrics detect no significant differences (see Fig. 11). We conclude that the periodicity patterns of the presence time `tmSess` and the number of clicks `cntCl` could be effectively applied to evaluate changes of the UI. We report the number of UI experiments passed by each metric by the *first* and the *second* digits of the *subscript* in the cells of Table 2.

Evaluation of server slowdown. We conduct 5 A/B experiments, where the server response time is artificially increased by $m\tau$ seconds[23], $m = 1, .., 5$. We find that only the normalized amplitudes $\mathcal{A}_{\mathcal{N}}$ of the presence time `tmSess` detected both of the two largest response time increments as significant changes (almost all $p_{\text{val}} < 10^{-2}$). These metrics outperform the baseline, see Fig. 12. We report the number of response time experiments passed by other metrics by the *second digit* of the *superscript* in the cells of Table 2.

Discussion. The results of all 25 A/B experiments for all 60 studied quality metrics are summarized in Table 2: the first number in a cell is the number of passed ($p_{\text{val}} > 0.05$) control experiments (5 in total) and the second one is the number of passed ($p_{\text{val}} \leq 0.05$) deterioration experiments (20 in total). The latter one is detailed by indexes in the cell: the digits in the superscripts report the numbers of passed ranking (3 in total) and response time (5 in total) experiments; the ones in the subscripts report the number of passed UI (7 in total) and light UI (5 in total) experiments.

In each row of Table 2 we chose the best metrics in the following way. Among quality metrics that pass all 5 control experiments, we highlighted in **boldface** the ones with the highest number of passed deterioration experiments. Among quality metrics that pass 4 control experiments, we also chose those with the highest number of passed deterioration experiments. In the cases, when this number exceeds that of the best metrics without fails on control experiments, we highlight the metric in blue color. The first column contains 4 baseline metrics (BS).

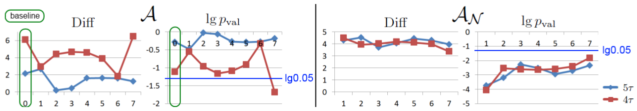

Figure 12: Results of 2 experiments with response time increasing for the presence time metric, for $\mathcal{A}$ and $\mathcal{A}_{\mathcal{N}}$ transformations.

The significant deterioration of the main web service component leads to strong changes in all metrics (all normalized amplitudes $\mathcal{A}_{\mathcal{N}}$ and almost all amplitudes $\mathcal{A}$ for the "swap-2" experiment, Fig. 9). But, when a change becomes less noticeable, different metrics show different sensitivity levels. Some studied metrics are sensitive to all changes in any component of the search engine (e.g., metrics derived from the presence time `tmSess`), the others specialize on catching differences in only one component (e.g., some metrics derived from the number of queries `cntQu` for "PersOff" experiment). If we allow using metrics that fail one of the control experiment, we obtain several quality metrics based on `cntQu`, which are more sensitive than the `cntQu`-based metrics without such fails (the last row in Table 2).

All metrics derived form the session count `cntSess` well pass the control A/A experiments and some of them catch both the ranking changes ($\mathcal{A}$ tend to grow, whereas $\mathcal{A}_{\mathcal{N}}$ tend to drop) and some UI changes (both $\mathcal{A}$ and $\mathcal{A}_{\mathcal{N}}$ tend to drop). The metrics derived from the presence time `tmSess` are most sensitive and almost all of them catch all types of studied changes with very high statistical significance level. The values Diff of the normalized amplitudes $\mathcal{A}_{\mathcal{N}}$(`tmSess`) for the response time changes are 5 times larger than the ones for the ranking changes. The metrics derived from the click count `cntCl` show the results similar to those for the presence time `tmSess` (except for the response time changes). This is naturally enough, since additional clicks lead to additional presence time. This assumption is supported by the results for the response time changes, where the changes in the presence time are caused by the engine efficiency and not by the user behavior. The metrics derived from the query count `cntQu` show their advantage on the experiment with personalization switching off, where one of its amplitudes and one of its normalized amplitudes outperform the baseline quality metrics. The users exposed to non-personalized results tend to submit more queries for some search tasks, what causes the changes in the periodicity amplitudes.

Finally, we conclude that *the baseline quality metrics are significantly outperformed by vast majority of the new periodicity metrics both in terms of significance level (p-value) and sensitivity to different search engine changes.* It is the answer to the **RQ3**.

8. CONCLUSIONS AND FUTURE WORK

In our work, we introduced *the periodicity pattern* (by means of the DFT) as an unified form for representation of user behavior periodicity. Namely, for a user, for 4-week pe-

---

[23] The factor τ is hidden for confidentiality reasons.

Table 2: The number of passed A/B experiments by each of the 60 quality metrics (see Section 7 for details).

$\mathbf{x}$	BS $\mathcal{A}(\mathbf{x})_0$	$\mathcal{A}(\mathbf{x})_1$	$\mathcal{A}(\mathbf{x})_2$	$\mathcal{A}(\mathbf{x})_3$	$\mathcal{A}(\mathbf{x})_4$	$\mathcal{A}(\mathbf{x})_5$	$\mathcal{A}(\mathbf{x})_6$	$\mathcal{A}(\mathbf{x})_7$	$\mathcal{A}_{\mathcal{N}}(\mathbf{x})_1$	$\mathcal{A}_{\mathcal{N}}(\mathbf{x})_2$	$\mathcal{A}_{\mathcal{N}}(\mathbf{x})_3$	$\mathcal{A}_{\mathcal{N}}(\mathbf{x})_4$	$\mathcal{A}_{\mathcal{N}}(\mathbf{x})_5$	$\mathcal{A}_{\mathcal{N}}(\mathbf{x})_6$	$\mathcal{A}_{\mathcal{N}}(\mathbf{x})_7$
`cntSess`	$5 \mid 3\,\left[\begin{smallmatrix}11\\10\end{smallmatrix}\right]$	$5 \mid 3\,\left[\begin{smallmatrix}10\\20\end{smallmatrix}\right]$	$5 \mid 1\,\left[\begin{smallmatrix}10\\00\end{smallmatrix}\right]$	$5 \mid 2\,\left[\begin{smallmatrix}20\\00\end{smallmatrix}\right]$	$5 \mid 1\,\left[\begin{smallmatrix}10\\00\end{smallmatrix}\right]$	$\mathbf{5} \mid \mathbf{4}\,\left[\begin{smallmatrix}20\\11\end{smallmatrix}\right]$	$5 \mid 3\,\left[\begin{smallmatrix}11\\01\end{smallmatrix}\right]$	$5 \mid 1\,\left[\begin{smallmatrix}10\\00\end{smallmatrix}\right]$	$\mathbf{5} \mid \mathbf{4}\,\left[\begin{smallmatrix}20\\20\end{smallmatrix}\right]$	$5 \mid 2\,\left[\begin{smallmatrix}20\\00\end{smallmatrix}\right]$	$5 \mid 3\,\left[\begin{smallmatrix}11\\10\end{smallmatrix}\right]$	$5 \mid 2\,\left[\begin{smallmatrix}11\\00\end{smallmatrix}\right]$	$5 \mid 1\,\left[\begin{smallmatrix}10\\00\end{smallmatrix}\right]$	$5 \mid 2\,\left[\begin{smallmatrix}10\\10\end{smallmatrix}\right]$	$5 \mid 2\,\left[\begin{smallmatrix}10\\10\end{smallmatrix}\right]$
`tmSess`	$5 \mid 3\,\left[\begin{smallmatrix}11\\10\end{smallmatrix}\right]$	$4 \mid 3\,\left[\begin{smallmatrix}10\\20\end{smallmatrix}\right]$	$5 \mid 2\,\left[\begin{smallmatrix}10\\10\end{smallmatrix}\right]$	$5 \mid 1\,\left[\begin{smallmatrix}10\\00\end{smallmatrix}\right]$	$5 \mid 4\,\left[\begin{smallmatrix}10\\30\end{smallmatrix}\right]$	$4 \mid 3\,\left[\begin{smallmatrix}10\\20\end{smallmatrix}\right]$	$5 \mid 3\,\left[\begin{smallmatrix}10\\20\end{smallmatrix}\right]$	$5 \mid 3\,\left[\begin{smallmatrix}11\\10\end{smallmatrix}\right]$	$5 \mid 10\,\left[\begin{smallmatrix}12\\34\end{smallmatrix}\right]$	$\mathbf{5} \mid \mathbf{12}\,\left[\begin{smallmatrix}12\\72\end{smallmatrix}\right]$	$5 \mid 11\,\left[\begin{smallmatrix}12\\62\end{smallmatrix}\right]$	$5 \mid 10\,\left[\begin{smallmatrix}12\\52\end{smallmatrix}\right]$	$5 \mid 7\,\left[\begin{smallmatrix}12\\31\end{smallmatrix}\right]$	$5 \mid 10\,\left[\begin{smallmatrix}12\\52\end{smallmatrix}\right]$	$4 \mid 10\,\left[\begin{smallmatrix}12\\61\end{smallmatrix}\right]$
`cntCl`	$5 \mid 2\,\left[\begin{smallmatrix}11\\00\end{smallmatrix}\right]$	$5 \mid 4\,\left[\begin{smallmatrix}10\\30\end{smallmatrix}\right]$	$5 \mid 2\,\left[\begin{smallmatrix}10\\10\end{smallmatrix}\right]$	$5 \mid 1\,\left[\begin{smallmatrix}10\\00\end{smallmatrix}\right]$	$5 \mid 2\,\left[\begin{smallmatrix}10\\10\end{smallmatrix}\right]$	$5 \mid 1\,\left[\begin{smallmatrix}10\\00\end{smallmatrix}\right]$	$5 \mid 2\,\left[\begin{smallmatrix}10\\10\end{smallmatrix}\right]$	$5 \mid 1\,\left[\begin{smallmatrix}10\\00\end{smallmatrix}\right]$	$5 \mid 8\,\left[\begin{smallmatrix}10\\61\end{smallmatrix}\right]$	$4 \mid 10\,\left[\begin{smallmatrix}11\\71\end{smallmatrix}\right]$	$4 \mid 10\,\left[\begin{smallmatrix}11\\71\end{smallmatrix}\right]$	$4 \mid 10\,\left[\begin{smallmatrix}11\\71\end{smallmatrix}\right]$	$\mathbf{5} \mid \mathbf{10}\,\left[\begin{smallmatrix}11\\71\end{smallmatrix}\right]$	$4 \mid 10\,\left[\begin{smallmatrix}11\\71\end{smallmatrix}\right]$	$4 \mid 9\,\left[\begin{smallmatrix}11\\70\end{smallmatrix}\right]$
`cntQu`	$5 \mid 1\,\left[\begin{smallmatrix}10\\00\end{smallmatrix}\right]$	$5 \mid 2\,\left[\begin{smallmatrix}00\\20\end{smallmatrix}\right]$	$\mathbf{5} \mid \mathbf{3}\,\left[\begin{smallmatrix}10\\20\end{smallmatrix}\right]$	$\mathbf{5} \mid \mathbf{3}\,\left[\begin{smallmatrix}10\\20\end{smallmatrix}\right]$	$5 \mid 0\,\left[\begin{smallmatrix}00\\00\end{smallmatrix}\right]$	$5 \mid 2\,\left[\begin{smallmatrix}10\\10\end{smallmatrix}\right]$	$5 \mid 1\,\left[\begin{smallmatrix}00\\10\end{smallmatrix}\right]$	$\mathbf{5} \mid \mathbf{3}\,\left[\begin{smallmatrix}10\\20\end{smallmatrix}\right]$	$4 \mid 4\,\left[\begin{smallmatrix}20\\20\end{smallmatrix}\right]$	$4 \mid 3\,\left[\begin{smallmatrix}11\\10\end{smallmatrix}\right]$	$4 \mid 2\,\left[\begin{smallmatrix}10\\10\end{smallmatrix}\right]$	$\mathbf{5} \mid \mathbf{3}\,\left[\begin{smallmatrix}11\\10\end{smallmatrix}\right]$	$4 \mid 3\,\left[\begin{smallmatrix}11\\10\end{smallmatrix}\right]$	$4 \mid 3\,\left[\begin{smallmatrix}11\\10\end{smallmatrix}\right]$	$4 \mid 1\,\left[\begin{smallmatrix}10\\00\end{smallmatrix}\right]$

riod, for 4 time series (the number of sessions, the presence time, the number queries and the number of clicks) we studied the amplitudes of the sine waves which compose the time series w.r.t. the DFT. Our basic and cluster analyses of user periodicity behavior show, that the periodicity patterns provide strong identification of several large user clusters with common coherent periodicity behavior.

Then, we studied how the periodicity patterns describe a long-term user behavior (during 20-week period). We found that the periodicity behavior, encoded in the periodicity patterns, is significantly more stable in long-term perspective than the user behavior data ranged by time. We find also that the users with definitive models of periodicity behavior stick to one periodicity model across months and the changes in their models are mostly caused by the changes in their life style (e.g., taking Christmas holidays). Thus, the periodicity patterns catch individual user traits that are mostly stable during a long time period.

We applied our periodicity patterns to a practical problem. We used the periodicity patterns as quality metrics (56 new scalar metrics in total) of search engine changes (such as the ranking algorithm, the UI, and the engine efficiency changes) by means of A/B tests. We found that the baseline quality metrics are significantly outperformed by vast majority of the new periodicity metrics both in terms of significance level (p-value) and sensitivity to different search engine changes.

Future work. We believe that our periodicity patterns will be of interest to researchers to be used in many practical problems. Our work is the first touch in this area, and in the future we can, first, extend the set of user engagement measures by investigating more sophisticated ones. Second, we can study the periodicity of user behavior across several web services. Third, we can also experiment with the size of the time window, e.g., study the periodicity patterns within a day. Finally, we can study the relationships between the periodicity user engagement metrics and other search engine quality metrics such as satisfaction, switching, etc.

9. REFERENCES

[1] E. Adar, J. Teevan, and S. T. Dumais. Large scale analysis of web revisitation patterns. In *CHI'2008*, pages 1197–1206, 2008.

[2] S. Chakraborty, F. Radlinski, M. Shokouhi, and P. Baecke. On correlation of absence time and search effectiveness. In *SIGIR'2014*, pages 1163–1166, 2014.

[3] A. Deng, Y. Xu, R. Kohavi, and T. Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *WSDM'2013*, pages 123–132, 2013.

[4] A. Diriye, R. White, G. Buscher, and S. Dumais. Leaving so soon?: understanding and predicting web search abandonment rationales. In *CIKM'2012*, pages 1025–1034, 2012.

[5] G. Dupret and M. Lalmas. Absence time and user engagement: evaluating ranking functions. In *WSDM'2013*, pages 173–182, 2013.

[6] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *SIGIR'2010*, pages 34–41, 2010.

[7] Q. Guo, R. W. White, Y. Zhang, B. Anderson, and S. T. Dumais. Why searchers switch: understanding and predicting engine switching rationales. In *SIGIR'2011*, pages 335–344, 2011.

[8] A. Hassan, X. Shi, N. Craswell, and B. Ramsey. Beyond clicks: query reformulation as a predictor of search satisfaction. In *CIKM'2013*, pages 2019–2028, 2013.

[9] A. Hassan, Y. Song, and L.-w. He. A task level metric for measuring web search satisfaction and its application on improving relevance estimation. In *CIKM'2011*, pages 125–134, 2011.

[10] A. Hassan and R. W. White. Personalized models of search satisfaction. In *CIKM'2013*, pages 2009–2018, 2013.

[11] V. Hu, M. Stone, J. Pedersen, and R. W. White. Effects of search success on search engine re-use. In *CIKM'2011*, pages 1841–1846. ACM, 2011.

[12] B. J. Jansen, A. Spink, and V. Kathuria. How to define searching sessions on web search engines. In *Advances in Web Mining and Web Usage Analysis*, pages 92–109. Springer, 2007.

[13] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu. Seven rules of thumb for web site experimenters. In *KDD'2014*, 2014.

[14] T. Kramár and M. Bieliková. Context of seasonality in web search. In *Advances in Information Retrieval*, pages 644–649. Springer, 2014.

[15] J. Lehmann, M. Lalmas, G. Dupret, and R. Baeza-Yates. Online multitasking and user engagement. In *CIKM'2013*, pages 519–528, 2013.

[16] J. Lehmann, M. Lalmas, E. Yom-Tov, and G. Dupret. Models of user engagement. In *User Modeling, Adaptation, and Personalization*, pages 164–175. Springer, 2012.

[17] K. Radinsky, K. Svore, S. Dumais, J. Teevan, A. Bocharov, and E. Horvitz. Modeling and predicting behavioral dynamics on the web. In *WWW*, 2012.

[18] K. Rodden, H. Hutchinson, and X. Fu. Measuring the user experience on a large scale: user-centered metrics for web applications. In *SIGCHI'2010*, pages 2395–2398, 2010.

[19] D. Savenkov, D. Lagun, and Q. Liu. Search engine switching detection based on user personal preferences and behavior patterns. In *SIGIR'2013*, pages 33–42, 2013.

[20] M. Shokouhi. Detecting seasonal queries by time-series analysis. In *SIGIR'2011*, pages 1171–1172, 2011.

[21] Y. Song, X. Shi, and X. Fu. Evaluating and predicting user engagement change with degraded search relevance. In *WWW'2013*, pages 1213–1224, 2013.

[22] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In *SIGMOD'2004*, pages 131–142, 2004.

[23] W. W.-S. Wei. *Time series analysis*. Addison-Wesley Redwood City, California, 1994.

[24] R. West, R. W. White, and E. Horvitz. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In *WWW'2013*, pages 1399–1410, 2013.

[25] R. W. White and S. T. Dumais. Characterizing and predicting search engine switching behavior. In *CIKM'2009*, pages 87–96, 2009.

[26] R. W. White, A. Kapoor, and S. T. Dumais. Modeling long-term search engine usage. In *User Modeling, Adaptation, and Personalization*, pages 28–39. Springer, 2010.

[27] Y. Zhang, B. J. Jansen, and A. Spink. Time series analysis of a web search engine transaction log. *Information Processing & Management*, 45(2):230–245, 2009.