9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Temporal Rhythms of Information Consumption on Wikipedia

Keywords: Wikipedia, Logs Analysis, Information seeking, Readers behavior, Daily Rhythms

Extended Abstract

Wikipedia is the largest encyclopedia ever created and one of the largest platforms for open knowledge. The English edition alone contains more than 6.5M articles, growing with a rate of 17k new articles each month[1]. Given this large pool of articles, Wikipedia covers a wide variety of topics and offers content to fulfill many diverse information needs. These differences in information needs bring readers to visit Wikipedia for reasons that vary from completing a school/work assignment to boredom. Previous studies [1] investigated how readers consume the content on Wikipedia. However, they overlooked a crucial dimension of information needs: the temporal daily rhythms. Although *time* is recognized in information science as a crucial contextual factor that drives human information seeking [2], it is often poorly investigated. The diurnal cycle is influenced by many aspects, including our circadian rhythm and environmental aspects that affect the information we need at different times. For example, during an average evening, we may be more inclined to look for information about a TV show than a math theorem. Naturally, given the substantial amount of time we spend browsing the Web, this tendency extends to online attention, including the content we consume on Wikipedia.

Thus, in this work, we characterize how information consumption in Wikipedia depends on the time of the day and how time interacts with other contextual properties, such as article topics and reader country. We complement previous studies on the consumption and popularity of Wikipedia content by –for the first time– analyzing the access log with timestamps converted into local time to remove the effect of different timezones.

Data. Our study relies on the access logs collected over four weeks –from March 1st to 28th, 2021– on the servers of the English edition of Wikipedia. These logs offer metadata such as the article loaded and the approximate geolocation. To prepare the data for this study, we select the requests relative to article views, which we anonymize by removing sensitive information as described in previous work [1]. Then, we convert all the timestamps into local time using the timezone information, and we retain 3.45B pageloads associated with 6.3M articles. We represent the number of pageloads of article a for each hour h of the day (averaged over 28 days) by $n_{a,h}$, i.e. each article is represented as a time series with $h = 0, \ldots, 23$.

Method. We define the normalized and timezone-corrected consumption pattern of all Wikipedia as the *baseline rhythm*, and it describes the expected access volume variation over time. We aim to obtain a temporal signature of each article by removing the baseline rhythm and focusing on how their consumption pattern diverges from this global average. We calculate the divergence by computing the per-hour ratio between each article's time series $n_{a,h}$ (additionally normalized so each time series sums to 1), and the baseline rhythm $N_h = \sum_a n_{a,h}$.

Fig. 1 summarizes these steps with two articles associated with different topics (STEM and MEDIA). Fig. 1a shows the shape of the curves after the normalization, and Fig. 1b shows their divergence from the baseline rhythm for the same articles.

[1] `https://en.wikipedia.org/wiki/Wikipedia:Statistics`

Summary of findings. To investigate the prototypical daily consumption patterns, we focus on two research questions: [RQ1] What are the typical shapes of Wikipedia consumption rhythms? [RQ2] Which factors influence the rhythms of Wikipedia consumption?

To approach RQ1, we extract the principal components of the matrix describing the normalized 24-dimension time series using SVD. We found that the first component models the day vs. night consumption behavior, capturing 39.1% of the variance. To explore if articles are organized into different prototypical consumption groups, we represent the access pattern with the value of their four principal components and use k-means to cluster the patterns. Interestingly, the articles cannot easily be separated into distinct, discrete groups. Manual inspection of the patterns reduced to two dimensions (UMAP) shows that the consumption behavior of the articles is distributed on a continuum (Fig. 2a), but with a tendency to group by topics (Fig. 2b).

To approach RQ2, we decompose each article's time series by country and access method. Then, we normalize each temporal pattern, remove the baseline rhythm, explode the time series into 24 samples with an explicit feature indicating the hour of the day, and complement with the relative binary vector representing the predicted article's topics obtained from ORES[2]. The obtained model fits the data with a R^2 of 0.181. Sorting the 24 coefficients representing the interactions allows us to characterize the temporal shape of each factor.

Fig. 3 shows the temporal shapes of the topics organized in the five top-level groups. The plot highlights how articles about STEM topics tend to receive more attention than average during the daytime and a visible reduction outside the typical working hours. On the other hand, articles about *Films*, *Television*, and *Biographies* have an inverted shape, with less consumption than average during the day and a substantial increase during the evening. Interestingly, the shapes of the temporal patterns suggest that content about *Video Games*, *Comics & Anime*, *Internet Culture*, *Military*, and *Society* are consumed by night-owl readers. Some of the shapes, especially the ones associated with STEM articles, show a reduction of attention around noon, suggesting that they might be affected by the lunch break when people's attention moves to other content types. This is corroborated by the fact that attention to articles about FOOD registers an increase during common meal times. Next, Fig. 4 shows the interaction coefficients of the countries with time. Some countries share similar daily patterns. Readers from the United States and Germany tend to consume Wikipedia more than average in the early morning, differently from readers from India and Ireland, that during the same hours, consume less content than average. Some countries, like Italy and Spain, reveal shared habits, such as reducing information consumption around noon, possibly associated with lunchtime.

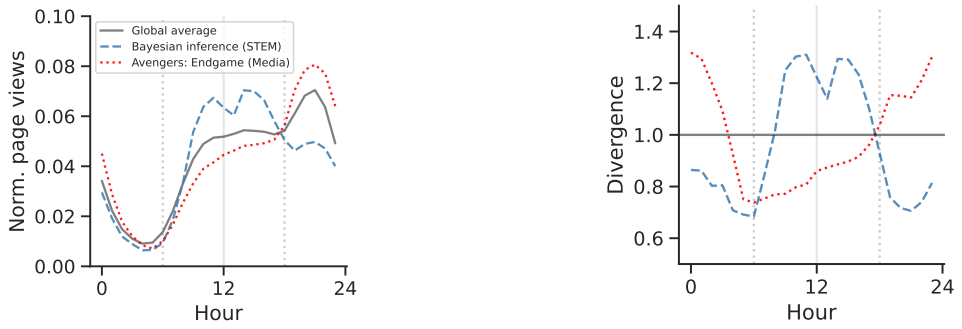
Conclusion. In this study, we have provided the first in-depth overview of how our daily rhythms affect how we consume information in one of the largest open knowledge platforms, Wikipedia. We hope our study provides a step towards operationalizing information consumption to better serve the needs of Wikipedia readers and web users in general.

References

[1] Tiziano Piccardi, Martin Gerlach, Akhil Arora, and Robert West. A large-scale characterization of how readers browse wikipedia. *ACM Trans. Web*, jan 2023.

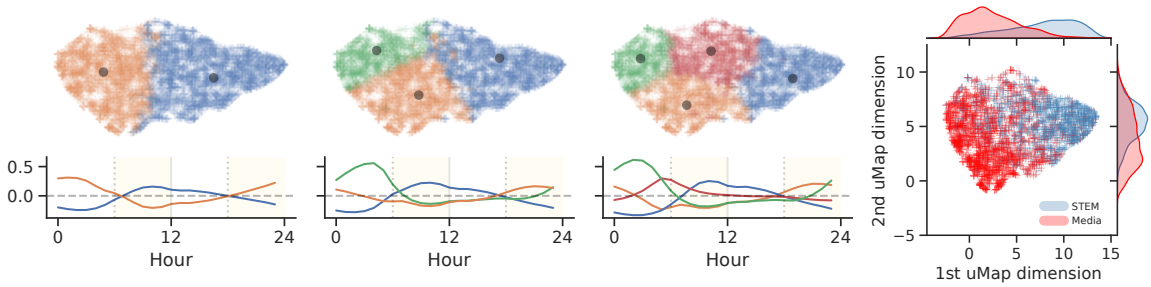
[2] Reijo Savolainen. Time as a context of information seeking. *Library & information science research*, 28(1):110–127, 2006.

[2] `https://www.mediawiki.org/wiki/ORES`



(a) Daily pattern (b) Divergence

Figure 1: Example of daily pattern of two articles about different topics – STEM dashed blue curve and Media dotted red curve. (a) Normalised daily time series ($n_{a,h}$). (b) Divergence from the normalised baseline rhythm ($D_{a,h}$)).



(a) k-means clusters (b) STEM vs. Media

Figure 2: (a) Top: UMAP projection of 10K articles. Clusters obtained with k-means for different values of $k = 2, 3, 4$. Bottom: Shape of the centroids reconstructed from their first 4 principal components. (b) Articles with color assigned based on their topics

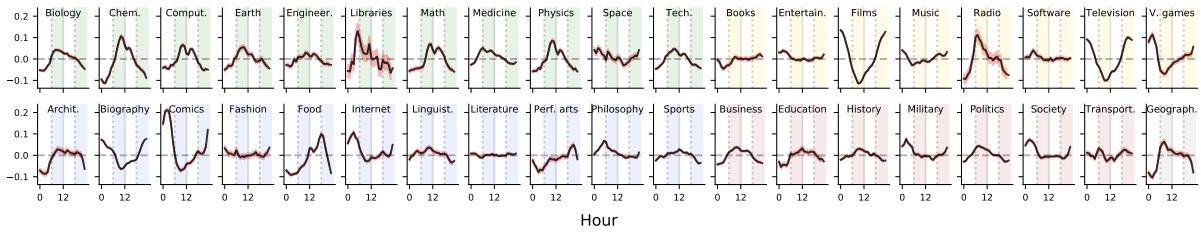


Figure 3: Coefficients of the interaction between topics and hours. Background colors represent the high-level classification: STEM (green), Media (yellow), Culture (blue), History & Society (red) and Geographical (grey).

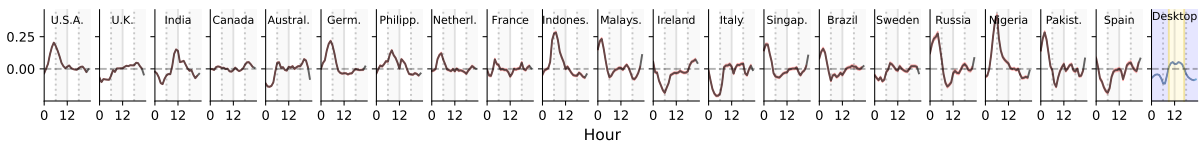


Figure 4: Coefficients of the interaction between country with hours (left, gray background, sorted by the most frequent origin of the requests), and device with hours (right, last plot). Yellow area for the desktop coefficients represents the typical working hour in western countries.