Seasonality Discovery of Large-Scale Wikipedia Page View Time Series Dataset

Xinyu Chen
Massachusetts Institute of Technology
Cambridge, USA
chenxy346@gmail.com

HanQin Cai
University of Central Florida
Orlando, USA
hqcai@ucf.edu

Lijun Ding
University of California, San Diego
La Jolla, USA
l2ding@ucsd.edu

Jinhua Zhao
Massachusetts Institute of Technology
Cambridge, USA
jinhua@mit.edu

Abstract

Large-scale Wikipedia page view observations represent the time series of human behaviors on digital platforms, showing behavioral rhythms and complicated patterns across daily and weekly cycles. This study crafts the frequently-viewed Wikipedia pages with the total amount of 3 million, which are selected according to the heavy-tailed distribution of page views. By leveraging the sparse autoregression model, the multi-cycle seasonality of Wikipedia page view time series can be clearly identified. The experiment demonstrates that the hourly page view time series of most frequently-viewed pages are less periodic than the relatively less viewed pages, namely, the page views of popular Wikipedia pages are less seasonal. The interpretable results delivered in this work could provide insights into understanding rhythms of human behaviors on digital platforms. In addition, the dataset used in this work is publicly available at https://doi.org/10.5281/zenodo.17070469.

CCS Concepts

• **Information systems** → *Information systems applications.*

Keywords

Wikipedia Page Views, Time Series Data, Data Mining, Seasonality, Autoregression, Sparsity

ACM Reference Format:
Xinyu Chen, HanQin Cai, Lijun Ding, and Jinhua Zhao. 2026. Seasonality Discovery of Large-Scale Wikipedia Page View Time Series Dataset. In *Proceedings of Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Wikipedia is one of the most visited website worldwide. Modeling Wikipedia page view time series is meaningful for understanding user behaviors [8], popularity of pages [10], external campaigns [5], and dynamic structures of web traffic [11]. Wikipedia page view data represent an important source for monitoring complicated behaviors and patterns underlying millions of Wikipedia pages. In literature, it verifies that the consumption habits of individual Wikipedia articles maintain strong diurnal regularities (i.e., temporal regularities) and reveal prototypical shapes of consumption patterns [11]. In addition to the temporal regularities, counterfactual predictions, i.e., quantifying the causal effect of effect by comparing the counterfactual predictions and actual page views, can be used to estimate how various external campaigns and events affect readership on Wikipedia [5].

Regarding a broad viewpoint of time series periodicity, temporal regularities such as weekly seasonality and multi-cycle autocorrelations underlying real-world time series are almost everywhere, including human activities (e.g., mobile phone usage [1]), urban human mobility [6, 7], and global climate variables (e.g., temperature and precipitation [7]). Substantial progress has been made for developing statistics and machine learning methods and utilizing them to quantify seasonality of real-world time series data. By using matrix factorization methods, time series data of human activities can be decomposed into low-rank factor matrices which could uncover daily, weekly, and seasonal rhythms and interpretable patterns [1]. Principal components of consumption habits of Wikipedia pages can be used to characterize remarkable diurnal regularities and the prototypical shapes of consumption patterns [11]. Recently, interpretable sparse autoregression methods have been developed for discovering time series periodicity from real-world systems [7]. These methods enable us to uncover the temporal regularities and inherent patterns simultaneously.

Although the analysis of temporal regularities is significant for identifying the patterns of web traffic, the literature ignored the simultaneous discovery of seasonality and inherent temporal patterns. The question also arises as how to formulate a comparable quantification in ultra-high dimensional time series data. This work contributes to the scientific community in the following ways:

- We provide an innovative data processing method by utilizing the heavy-tailed distributions of Wikipedia page view time series. The data samples show a shrinkage of zero-inflated and biased observations, maintaining a small fraction of Wikipedia pages which dominant the total traffic. In particular, the released dataset can be used as an important

WWW '26 Companion, Dubai, United Arab Emirates

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

benchmark and testbed for the data mining and machine learning tasks.
- We present an interpretable sparse autoregression method for understanding the temporal regularities and multi-cycle seasonality of high-dimensional page view time series. The method can be used to discover the daily and weekly seasonality of Wikipedia page views of different levels. The results demonstrate that the page view time series of popular and frequently-viewed Wikipedia pages are less seasonal than less viewed pages.

2 Dataset

The Wikipedia page view data portal includes long-term hourly page view observations since May 2015 across more than 60 million Wikipedia pages.[1] As shown in Figure 1A, the empirical distribution of Wikipedia page views that uses logarithmic scales on both the horizontal and vertical axes at two certain hours on January 1, 2024 demonstrates heavy tails. That means that only a small fraction of Wikipedia pages have been frequently-viewed. By observing the time series of the number of pages and page views from January 1, 2024 to January 7, 2024 in Figure 1B, they demonstrate strong periodic patterns across different days. Although the pages whose number of page views greater than 1 (i.e., page view $\geq$ 2) are only a small portion of total Wikipedia pages, their page views are dominant as shown in Figure 1C. Thus, it motivates us to preprocess the page view data and extract the frequently-viewed pages. One remarkable advantage by doing so is mitigating the impact of zero-inflated and biased observations when learning a certain model.

To get the time series from the page view observations, the first task is data alignment across different Wikipedia pages in the time domain. In the original page view observations, the observed pages which have been viewed at least once are quite different between two hours. For example, Figure 1B shows that the total number of pages are changing over time, presenting the predictable ebbs and flows of web traffic data. The number of pages in peak hours is significantly greater than off-peak hours. By selecting the Wikipedia pages that have been viewed at least 10 times in each day, we can roughly obtain 5-6 million unique pages. We then build the sample time series of hourly page views for each day. Of the source data in January 2024, we identify the intersection set of pages across 31 days, referring to 3,031,046 unique Wikipedia pages. Although the selected 3 million pages of January 2024 are less than 5% in total Wikipedia pages, they have 12.88 billion total views and dominate the total page views in 72% across the whole month of January 2024. Figure 1D shows the total number of page views of each hour in January 2024, demonstrating a remarkable weekly periodicity where weekends usually have higher page views during peak hours than weekdays.

3 Approach

For any page view time series $\boldsymbol{x}_i=(x_{i,1},x_{i,2},\cdots,x_{i,T})^\top\in\mathbb{R}^T$ of the page $i\in\{1,2,\ldots,n\}$ (i.e., n pages in total) with an hourly time resolution, we divide the page view data into $\Gamma\in\mathbb{Z}^+$ categories according to the total number of page views in January 2024. Among these 3 million Wikipedia pages, we select the pages that have been viewed between 10^2 and 10^3 times in the whole month of January as the first category $\gamma=1$, denoted by $\mathcal{O}(10^2)$, consisting of 0.78 million pages. The second category $\gamma=2$ includes 2.04 million pages that have been viewed between 10^3 and 10^4 times. The third category $\gamma=3$ refers to the most frequently-viewed pages that have the number of page views between 10^4 and 10^5, consisting of 0.20 million pages. The preliminary clustering of the Wikipedia pages allows one to analyze the patterns of page view time series.

As a result, we have time series data $\boldsymbol{x}_{i,\gamma}\in\mathbb{R}^T$ for any page $i\in\{1,2,\ldots,n_\gamma\}$ in the category $\gamma\in\{1,2,\ldots,\Gamma\}$ (i.e., $\Gamma=3$ in this case), where each category has n_γ pages. This work aims to discover the multi-cycle seasonality of page view time series in these categories by using sparse autoregression proposed in [7]. The sparse autoregression is an efficient interpretable machine learning method for quantifying periodic patterns of real-world time series data with inherent periodic cycles. On the top of time series autoregression, the sparse autoregression assumes both sparsity and non-negativity of auto-correlations for automatically identifying dominant auto-correlations such as daily and weekly cycles as interpretable results. To learn the sparse auto-correlations that represent the seasonality of these page view time series, we rewrite the optimization problem of sparse autoregression as follows,

$$\begin{aligned}\min_{\{\boldsymbol{w}_\gamma\}_{\gamma=1}^{\Gamma},\boldsymbol{z}}\;&\sum_{\gamma=1}^{\Gamma}\sum_{i=1}^{n_\gamma}\|\boldsymbol{y}_{i,\gamma}-\boldsymbol{A}_{i,\gamma}\boldsymbol{w}_\gamma\|_2^2\\ \text{s.t. }&0\leq\boldsymbol{w}_\gamma\leq\mathcal{M}\cdot\boldsymbol{z},\,\forall\gamma\in\{1,2,\ldots,\Gamma\},\\ &\sum_{k=1}^{d}z_k\leq\tau,\\ &z_k\in\{0,1\},\,\forall k\in\{1,2,\ldots,d\},\end{aligned}\tag{1}$$

where $d\in\mathbb{Z}^+$ is the order of sparse autoregression. The sparsity level is denoted by $\tau\in\mathbb{Z}^+$, which is far smaller than the order d. The decision variable $\boldsymbol{z}\in\mathbb{R}^d$ is a binary vector, which is constrained to be τ-sparse. $\mathcal{M}\in\mathbb{R}^+$ is a sufficiently large constant. In the objective function, $\|\cdot\|_2$ denotes the ℓ_2-norm of any vector. The above-mentioned optimization differs from [7] in the way that: For each category $\gamma\in[\Gamma]$, the model only uses a sparse coefficient vector to express the dominant auto-correlations in a sequence of time series. Thus, the seasonality of each category represents the overall periodic pattern of thousands or millions of page view time series.

In practice, time series autoregression can be easily reformulated as a linear regression. In Eq. (1), the data pairs $\{\boldsymbol{y}_{i,\gamma},\boldsymbol{A}_{i,\gamma}\},\forall i\in\{1,2,\ldots,n_\gamma\},\gamma\in\{1,2,\ldots,\Gamma\}$ are constructed by the page view time series, i.e.,

$$\boldsymbol{y}_{i,\gamma}=\begin{bmatrix}x_{i,\gamma,d+1}\\x_{i,\gamma,d+2}\\\vdots\\x_{i,\gamma,T}\end{bmatrix},\quad\boldsymbol{A}_{i,\gamma}=\begin{bmatrix}x_{i,\gamma,d}&x_{i,\gamma,d-1}&\cdots&x_{i,\gamma,1}\\x_{i,\gamma,d+1}&x_{i,\gamma,d}&\cdots&x_{i,\gamma,2}\\\vdots&\vdots&\ddots&\vdots\\x_{i,\gamma,T-1}&x_{i,\gamma,T-2}&\cdots&x_{i,\gamma,T-d}\end{bmatrix},$$

where $x_{i,\gamma,t}$ is the t-th data point of the time series $\boldsymbol{x}_{i,\gamma}\in\mathbb{R}^T$.

[1] https://dumps.wikimedia.org/other/pageviews/readme.html

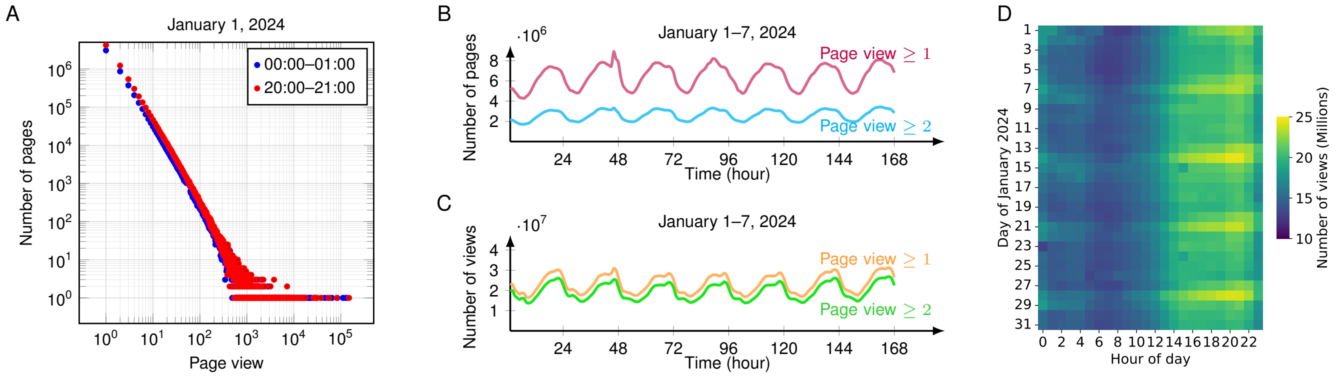

Figure 1: Empirical demonstration of the Wikipedia page view time series dataset. (A) Log-log plot of heavy-tailed distributions of page view data on January 1, 2024. The number of pages at 00:00 is 5.18 million, while 7.26 million at 20:00. Here, 60% pages have been viewed only once in the whole hour. (B) Hourly time series of number of pages that have been viewed. The data from January 1, 2024 to January 7, 2024 are selected for demonstration. In the panel, the notions "page view ≥ 1" and "page view ≥ 2" correspond to the page have been viewed at least once and twice, respectively. (C) Hourly time series of number of views. Although the number of pages that have been viewed at least twice is not a large portion (see Panel B), the difference of the number of views between the "page view ≥ 1" and the "page view ≥ 2" is marginal. (D) Hourly time series of number of views on the 3-million page data in January 2024. These views are up to 72% of the total Wikipedia page views.

According to the property of matrix trace, the objective function of this optimization problem can be reformulated as

$$\begin{aligned} f(\boldsymbol{w}_\gamma) &= \sum_{i=1}^{n_\gamma} \|\boldsymbol{y}_{i,\gamma} - \boldsymbol{A}_{i,\gamma}\boldsymbol{w}_\gamma\|_2^2 \\ &= \operatorname{tr}\Big(\boldsymbol{w}_\gamma\boldsymbol{w}_\gamma^\top \sum_{i=1}^{n_\gamma} \boldsymbol{A}_{i,\gamma}^\top \boldsymbol{A}_{i,\gamma}\Big) - 2\boldsymbol{w}_\gamma^\top \sum_{i=1}^{n_\gamma} \boldsymbol{A}_{i,\gamma}^\top \boldsymbol{y}_{i,\gamma} + C, \end{aligned} \tag{2}$$

where the operator $\operatorname{tr}(\cdot)$ denotes matrix trace. The last term C is the constant. By computing

$$\boldsymbol{P}_\gamma \triangleq \sum_{i=1}^{n_\gamma} \boldsymbol{A}_{i,\gamma}^\top \boldsymbol{A}_{i,\gamma} \in \mathbb{R}^{d\times d}, \quad \boldsymbol{q}_\gamma \triangleq \sum_{i=1}^{n_\gamma} \boldsymbol{A}_{i,\gamma}^\top \boldsymbol{y}_{i,\gamma} \in \mathbb{R}^{d},$$

from the data pairs in advance, we can build the following optimization problem:

$$\begin{aligned} \min_{\{\boldsymbol{w}_\gamma\}_{\gamma=1}^{\Gamma}, \boldsymbol{z}} \; & \sum_{\gamma=1}^{\Gamma} \big(\operatorname{tr}(\boldsymbol{w}_\gamma\boldsymbol{w}_\gamma^\top \boldsymbol{P}_\gamma) - 2\boldsymbol{w}_\gamma^\top \boldsymbol{q}_\gamma\big) \\ \text{s.t. } & 0 \leq \boldsymbol{w}_\gamma \leq \mathcal{M}\cdot \boldsymbol{z}, \; \forall \gamma \in \{1,2,\ldots,\Gamma\}, \\ & \sum_{k=1}^{d} z_k \leq \tau, \\ & z_k \in \{0,1\}, \; \forall k \in \{1,2,\ldots,d\}, \end{aligned} \tag{3}$$

which can be solved by mixed-integer optimization algorithms [2–4, 7]. Since the decision variable vectors $\boldsymbol{w}_\gamma, \gamma \in \{1,2,\ldots,\Gamma\}$ have the same sparsity constraint, these optimization results are both interpretable and comparable. Notably, the global sparsity of binary decision variable vector $\boldsymbol{z}$ is denoted by $\Omega = \operatorname{supp}(\boldsymbol{z})$, i.e., index set of the nonzero entries in $\boldsymbol{z}$.

4 Seasonality Exploration

We conduct extensive experiments for examining the sparse autoregression model on the Wikipedia page view time series dataset. The dataset is organized as a three-dimensional array, including Wikipedia page, category, and hourly time step dimensions. In the first category $\gamma = 1$, the page view time series are represented as a matrix of 0.78 million rows and 744 hours in the whole month of January. The time series matrices of the second and third categories have 2.04 million and 0.20 million rows, respectively. Although the matrices of three categories have different numbers of rows, one can compute $\boldsymbol{P}_\gamma \in \mathbb{R}^{d\times d}$ and $\boldsymbol{q}_r \in \mathbb{R}^{d}$ in advance, leading to the simplified objective function in Eq. (3). By using the mixed-integer programming algorithm in CPLEX [9], the sparse vectors $\{\boldsymbol{w}_\gamma\}_{\gamma\in[3]}$ are expected to be interpretable.

By setting the model's sparsity levels as $\tau = 8, 10$ and the order as $d = 168$, referring to a weekly cycle, one can obtain the positive auto-correlations (or coefficients) as shown in Figure 2. Under sparsity level $\tau = 8$, the support set is optimized as $\Omega = \{1,2,4,9,22,24,48,168\}$ (i.e., cardinality of this set is 8) where indices $\{24,48,168\}$ correspond to the daily, bi-daily, and weekly cycles, respectively. Therefore, the coefficients at index 24 represent the strength of daily seasonality. Accordingly, we can quantify the bi-daily and weekly seasonality by the coefficients at indices 48 and 168, respectively. By contrast, setting the sparsity level as $\tau = 10$ leads to the support set $\Omega = \{1,2,3,6,10,22,24,47,96,168\}$, in which indices $\{24,96,168\}$ correspond to the daily, four-day, and weekly cycles, respectively.

In these coefficients with the sparsity level $\tau = 8$ across three different page categories, one can find that the page view seasonality of frequently-viewed pages (i.e., monthly page view in $[10^4, 10^5)$) is lower than less-viewed pages (e.g., monthly page view in $[10^2, 10^3)$).

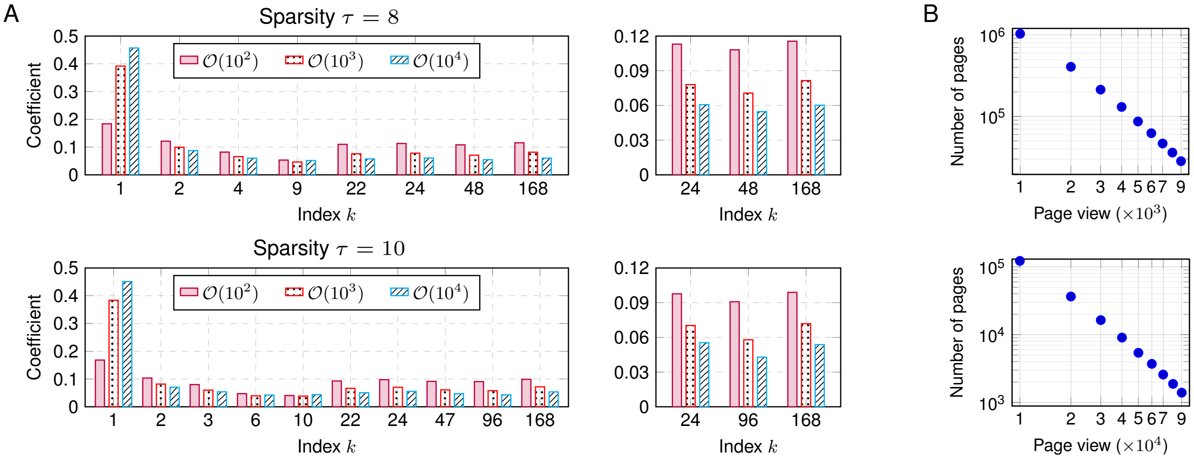

Figure 2: Seasonality analysis of Wikipedia page view time series with the sparse autoregression model. (A) Auto-correlations (i.e., autoregressive coefficients) on the optimized index set that controlled by the sparsity levels $\tau = 8, 10$, respectively. Sample page view time series are selected according to three categories of the number of page views in the whole month of January 2024, i.e., $[10^2, 10^3)$, $[10^3, 10^4)$, and $[10^4, 10^5)$, which are simplified as $O(10^2)$, $O(10^3)$, and $O(10^4)$ in the plot, respectively. The page numbers in these three categories are 0.78 million, 2.04 million, and 0.20 million, respectively. (B) Log-log plot of distributions of page view data. The top panel shows 9 categories with total number of page views ranging from 1,000 to 9,999 in January 2024. The bottom panel shows 9 categories with total number of page views ranging from 10,000 to 99,999.

To test the sensitivity of these results, we can compare the coefficients between sparsity levels $\tau = 8$ and 10 in the model. Although the optimized multi-cycle seasonality indices $\{24, 96, 168\}$ of the sparsity level $\tau = 10$ are a little different from the sparsity level $\tau = 8$, one can also see that the time series of frequently-view pages are less seasonal than the less-viewed pages. These findings can explain that the frequently-viewed pages might be venerable to special events and anomalies on digital platforms such as Wikipedia.

5 Conclusion

In this work, we creatively used the large-scale Wikipedia page view observations and empowered the analysis of page view time series seasonality. In particular, we built an hourly page view time series dataset according to the heavy-tailed distribution of page view observations. The pages that have been viewed at least 10 times in each day are selected, and consequently, the dataset has 3 million pages and they dominated more than 70% total page views on Wikipedia. On the 3 million time series, we analyzed the multi-cycle seasonality of page views with different levels of monthly page views by using the sparse autoregression model. The results demonstrate that the pages with high page views are less periodically viewed than the pages with lower page views. Both dataset and analysis throughout this work are insightful for understanding periodic human behaviors and rhythms on digital platforms. In the future, the large-scale Wikipedia page view time series data in this work could be utilized as a benchmark for time series seasonality quantification.

Acknowledgments

We acknowledge the Mens, Manus, and Machina (M3S) which is an interdisciplinary research group (IRG) of the Singapore MIT Alliance for Research and Technology (SMART) center.

References

[1] Talayeh Aledavood, Ilkka Kivimäki, Sune Lehmann, and Jari Saramäki. 2022. Quantifying daily rhythms with non-negative matrix factorization applied to mobile phone data. *Scientific reports* 12, 1 (2022), 5544.

[2] Dimitris Bertsimas, Vassilis Digalakis Jr, Michael Lingzhi Li, and Omar Skali Lami. 2024. Slowly varying regression under sparsity. *Operations Research* (2024).

[3] Dimitris Bertsimas, Angela King, and Rahul Mazumder. 2016. Best subset selection via a modern optimization lens. (2016).

[4] Dimitris Bertsimas and Bart Van Parys. 2020. Sparse high-dimensional regression. *The Annals of Statistics* 48, 1 (2020), 300–323.

[5] Xiaoxi Chelsy Xie, Isaac Johnson, and Anne Gomez. 2019. Detecting and gauging impact on Wikipedia page views. In *Companion Proceedings of The 2019 World Wide Web Conference*. 1254–1261.

[6] Xinyu Chen, HanQin Cai, Fuqiang Liu, and Jinhua Zhao. 2025. Correlating time series with interpretable convolutional kernels. *IEEE Transactions on Knowledge and Data Engineering* (2025).

[7] Xinyu Chen, Vassilis Digalakis Jr, Lijun Ding, Dingyi Zhuang, and Jinhua Zhao. 2025. Interpretable Time Series Autoregression for Periodicity Quantification. *arXiv preprint arXiv:2506.22895* (2025).

[8] Dimitar Dimitrov, Florian Lemmerich, Fabian FlÃ, Markus Strohmaier, et al. 2019. Different topic, different traffic: How search and navigation interplay on wikipedia. *The Journal of Web Science* 6 (2019).

[9] IBM ILOG CPLEX. 2015. CPLEX Optimization Studio CPLEX User's Manual, Version 12 Release 6. https://www.engineering.iastate.edu/~jdm/ee458/CPLEX-UsersManual2015.pdf.

[10] Innocensia Owuor, Hartwig H Hochmair, and Gernot Paulus. 2023. Use of social media data, online reviews and wikipedia page views to measure visitation patterns of outdoor attractions. *Journal of Outdoor Recreation and Tourism* 44 (2023), 100681.

[11] Tiziano Piccardi, Martin Gerlach, and Robert West. 2024. Curious rhythms: Temporal regularities of wikipedia consumption. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 1249–1261.