Using Signal processing to Investigate User Web Search Behaviour on Topics of Interest with Multiple Periodicities

Jivashi Nagar and Hussein Suleman

University of Cape Town, South Africa
(Jnagar, Hussein) @cs.uct.ac.za

Abstract. A Web user searches for multiple topics of interest on the Web on a regular basis. The topics may be searched periodically at a particular time or at different times, creating temporal patterns with different periodicities in the search history. To improve a user's Web search experience, the multiple periodicities of the topics of interest of a user can be exploited. This study proposes to find multiple periodicities of a user's topic of interest through signal processing. It is found that Fast Fourier Transform can be used to find multiple periodicities of a topic as well as to predict the temporal pattern of the topics with accuracy and error depending on the training data size.

Keywords: Web search, Temporal patterns, FFT, Multiple periodicities.

1 Introduction

The Web is extensively searched for various topics. Most of the topics searched by the Web-users are time-sensitive and are searched regularly [1–3]. But, different people have different topics and times preferences. For example, people who watch sports may issue more queries on Saturdays than on weekdays. Students are likely to issue more school-oriented questions in the evenings on weekdays when they are completing their homework. Accountants may issue queries on tax law that they will only submit during work hours from Monday to Friday. Users also have different interests at different times: the accountant could also watch sport. Thus to better model the user, we need to know their topics of interest and their periodic behaviour for each topic. Periodic behaviour can be complex, as it can be a simple pattern (every morning) or a combination of patterns (every morning on weekdays and all day on weekends).

Providing the search results for a query while ignoring the time of search, may not prove useful every time. The intent of a user associated with a query depends on the time of the search. This intent is referred to as the temporal intent. Understanding the temporal intent of a user can be utilized to provide relevant results [4].

Searching for specific topics at particular times makes the user's search history a temporal dataset with topics searched having different temporal patterns. For instance, a topic in a user's search history may have a single periodicity pattern or a multiple periodicity pattern. A temporal pattern is a single periodicity pattern if it has one period which means that the topic is searched at a particular time regularly. A topic has a multiple periodicity pattern if it has multiple periods. For example, a topic which is searched at different times on weekdays and weekends has multiple periods. In general, to find the temporal patterns in a temporal dataset, a single periodicity is defined and all the patterns with a required threshold periodicity occurring after a particular period are considered as temporal patterns. Thus, a temporal pattern is easy to retrieve and predict if it has a single periodicity. But, in the case of Web search history, there may be temporal patterns with multiple periodicities.

Fig. 1 shows a temporal pattern with multiple periodicities. The pattern repeats after a regular hourly period on alternate days. So, on the day, when it occurs, the subsequent occurrences are a few hours apart except for the case of its last occurrence, where the next occurrence is a day apart.

This study has proposed the use of Fast Fourier Transform (FFT) to find the multiple periods of a temporal pattern. As search logs from commercial Web search engines don't have open access, synthetic data has been created for the purpose of this study. The temporal pattern of a topic is treated as a signal and analysed using Fast Fourier Transform.

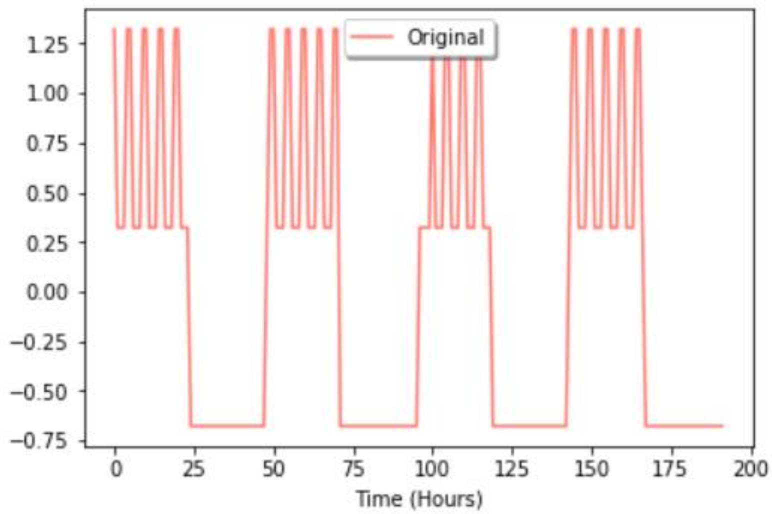

Fig. 1. A signal representing a user's topic of interest with multiple periodicities

2 Related work

The main aim of this work is to study the use of Fast Fourier Transform (FFT) in finding the multiple periods of topics searched by a user and to analyse the accuracy of prediction to ultimately improve the Web search experience. This section is divided into two sub-sections. The first sub-section briefly shows the other related works that have been done to provide personalized and relevant search results. The second sub-section discusses signal processing and its benefits studied in different areas.

2.1 Improving Web Search Experience

User Web search history has been studied widely to understand user search behaviour [5]. To predict the nature of the information need of a user, it is essential to study the past browsing behaviour of the user [6, 7]. The location has also been studied as a feature to identify the user intent associated with a query [8]. It is found that using location as a feature is not equally beneficial to all kinds of queries. Time has also been studied as an influential factor for personalization in various aspects. A study by Liu et al [9] proposed a model to identify the temporal information of a user's Web search using dwell and click sequence on search results. Item-level dwell time has also been capitalized as a measure of item relevance to a user [10]. According to a study by Halvey et al [11], users have specific needs at different times of the day and the week. They predicted user's mobile web navigation pattern using Markov models based on the automated method they developed to segment log data by time periods. Metzler et al [12] exploited the implicit temporal intent associated with search queries to provide relevant results. Vu et al [13] also modelled time in re-ranking the search results to deliver more relevant search results. According to this study, recent documents may be of more relevance to the user. Song and Guo have considered the time of search in their work to improve recommendation quality [3]. They generated query level, task-level and user-level features to capture the characteristics of task repetition and used neural networks to recommend better results. Similar to this work, the aim of our research is also to improve the user search experience by providing better search results using the time of the search. It is not always necessary that a user searches for a topic at a specific hour every day. It may be on alternate days or only on specific days. Our work also addresses the issue of multiple periodicities in a temporal pattern using signal processing. A signal or a pattern with multiple periodicities can be thought of a combination of different sub-signals or sub-patterns that have their own periodicities. For instance, a signal which shows that a topic is searched every morning at 8 am on weekdays and at 2 pm on weekends.

2.2 Fast Fourier Transform

Signal processing is extensively used in different areas like Electronics and communication, Biology, Astronomy, Time-series analysis and so on. FFT has found use in studying the quasi-periodic signals [14, 15] in radiation. To study the climate data in the frequency domain, the work by [16] used FFT. To predict seizures, a model has been proposed based on classifiers like Random Forest and SVM (Support Vector Machine) with FFT [17]. The study [18] proposed a model based on FFT with machine learning to advise if a patient needs a body check-up on the basis of past medical data of the patient. To reduce the computation and storage complexity of Deep Neural Networks (DNN), FFT has found significant use [19]. Multivariate time series classification has also been done with FFT [20]. Alexey et al [21] used FFT and spectrum analysis to find the periodicity patterns of user's engagement with search engines and to evaluate search quality. All of these results predict user behaviour based on periodic patterns.

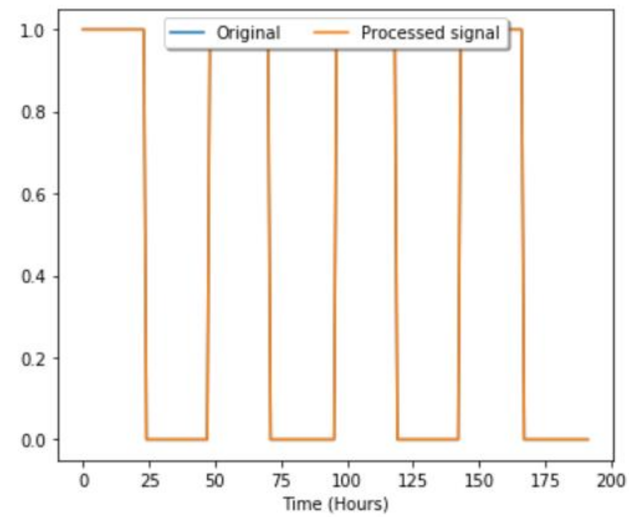

Fig. 2. Comparing the original pure signal with the processed signal

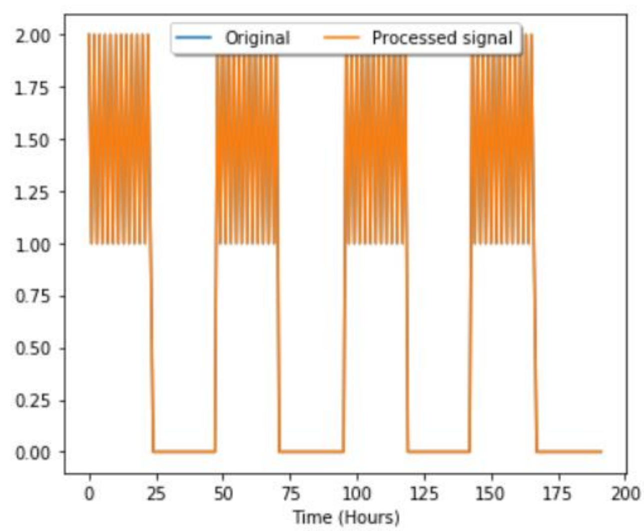

Fig. 3. Comparing the original mixed signal with the processed signal

3 Methodology

The experimental methodology consists of creating the sample signals, applying Fast Fourier Transform (FFT) and Inverse Fast Fourier Transform (IFFT), preparing a model for extrapolation, followed by comparing the results.

3.1 Creating signals

Fifteen main synthetic signals were created for the purpose of this study. They are composed of different periodicities. Each signal represents a sinusoidal wave. The length of each signal is 192 hours (8 days). Some of the signals are regarded as pure signals. For convenience, they are named as S1, S2, S3, S4, S5, and S6. In Signal S1, the pattern repeats after a regular interval of five hours on alternate days, while in Signal S2, the pattern repeats at each alternate hour on alternate days. Signal S3 represents a pattern that repeats at every hour on alternate days. Signal S4 shows a weekend pattern where a topic is searched only on Saturdays and Sundays. Signal S5 represents a signal where a topic is searched after 5 hours every day while Signal S6 shows that a search is made at alternate hours every day. Some signals are created by mixing some of the pure signals. Signals S12, S13, S23, S123, S35, S36, S45 and S46 are regarded as "mixed signals" as they are created by combining the "pure signals". These signals show the combination of different periodicities of their component pure signals. Thus, these signals represent multiple periodicities. For example, Signal S13 as shown in Fig. 1 is a combination of Signal S1 and Signal S3. It becomes clear from Figure1 that this signal is a combination of two waves or signals with different amplitudes and frequencies. The component signals of various frequencies can be filtered out using FFT.

Signal S0 also has a pattern but is burst with some random noise. It represents a topic that is often searched periodically but may be due to some other activities or engagements, the topic is not searched for a few regular periods. For example, a user searches about political news every morning but, when he goes for holidays, he doesn't search about them. Thus, noise is created in the periodic signal.

3.2 Applying FFT and IFFT

Fast Fourier Transform (FFT) is an algorithm to efficiently convert the Time-domain data to the Frequency-domain. In this study, FFT is applied to the signals to analyse the periodicities of a signal, as shown in Fig. 1. The FFT values are then converted back to the time-domain by applying the Inverse Fast Fourier Transform (IFFT). This allows for extrapolation to future time periods. IFFT produces the same values in time-domain. It is observed that after applying IFFT to the frequency domain values, the original time-domain signal is reproduced. Thus, FFT and IFFT can potentially be used to predict or extrapolate the original signal. The graphs in Fig. 2 and Fig. 3 are showing the comparison of the original signal and the processed signal, after applying IFFT to FFT values of Signal S3 and Signal S23. As it is apparent from these figures that both the original and the processed signals are overlapping. This means that using IFFT, the same original signal irrespective of the number of inherent periodicities, can be recreated. All these conversions and analysis are done in Python 3.6 using Numpy and Scipy libraries along with other required libraries for data processing and visualization.

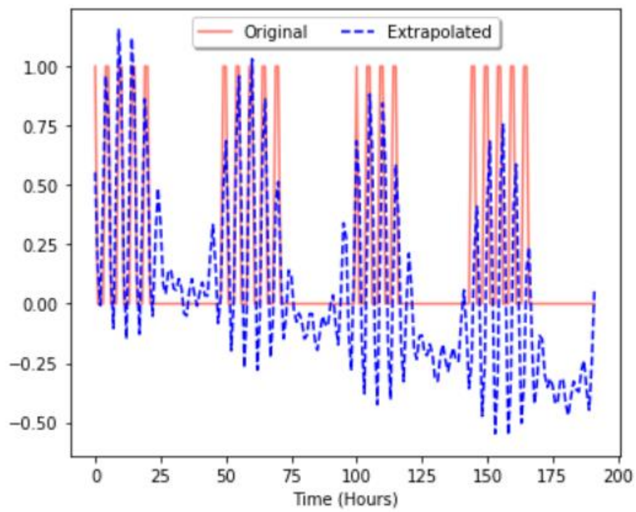

Fig. 4. A pure signal with its respective extrapolated signal with Time batch 0-96

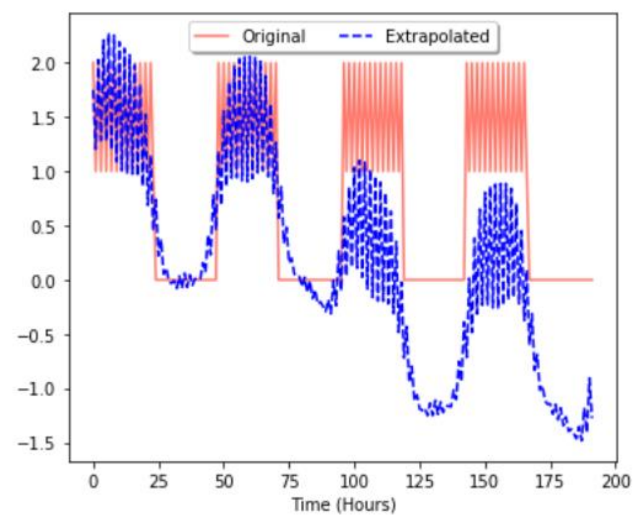

Fig. 5. A mixed-signal with its respective extrapolated signal with Time batch 0-96

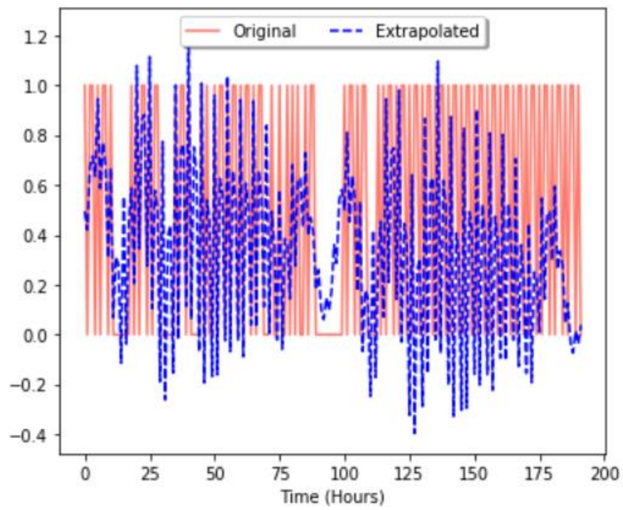

Fig. 6. A noisy signal with its respective extrapolated signal with Time batch 0-96

3.3 Extrapolation and Comparison

Each of the original signals is divided into Time batches, ranging from 0-24 hours to 0-192 hours, producing 8 segments of the original signal. A prediction model is trained on a segment and then extrapolated for the rest of the length of the original signal. For example, if the prediction model is trained on a segment of Time batch 0-48 hours, it will predict the signal from 49 hours to 192 hours.

The following steps explain the process of extrapolation:

1. Different segments of the original signal are created for different Time batches.
2. Size of Prediction signal is calculated by taking the difference between a segment of the original signal and the original signal.
3. The segment is cleaned of any constant trend. De-trending a signal means the removal of any linear trends in the signal to allow the identification of potential cyclic patterns in the signal.
4. The de-trended segment of the signal is converted to Frequency domain by applying FFT on it.
5. The magnitude is calculated by taking the absolute value of FFT values.
6. The phase is calculated by taking the angle of FFT values
7. The magnitude and the phase values are processed to construct the signal back in the time domain for the length of the original signal. This produces an extrapolated signal

Fig. 4, Fig. 5, and Fig. 6 show the extrapolated signal trained with a segment of a signal of Time batch 0-96 hours. Fig. 4 represents the extrapolated signal of the original signal Signal S1, Fig. 5 shows the extrapolated signal of the original signal Signal S36, and Fig. 6 represents the extrapolated signal of the original signal Signal S0.

To find the percentage accuracy and error of the extrapolated signals, the difference between rounded and absolute values of the extrapolated signal with each corresponding value of the original signal is calculated. Fig. 8, Fig. 9, Fig. 10, and Fig. 11 show the percentage of the accuracy and the error values of Signals S0, S1, S12, and S123 respectively with different Time batches.

4 Results and Discussion

It is observed that FFT is able to find multiple periods in a single signal. It can be used efficiently to filter the individual signals of a single period from a mixed-signal. The IFFT can produce the signal from the frequency domain to the time domain, similar to the original signal. In the case of extrapolation, with respect to different training Time batches, it is seen that more training data corresponds to more accuracy for all the different kinds of signals. As shown in Fig. 8. Fig. 9, Fig. 10, and Fig. 11 in general, the accuracy is increasing with an increase in the training data, except at some points. For Signal S1, the percentage of accuracy is higher than that of the error with most of the training Time batches. Similar is the case with Signal S12. Surprisingly, the Signal S0 that has noise incorporated within a regular pattern, shows high accuracy with every training Time batch. Signal S123 which is a combination of three signals of different periodicities, gives poor accuracy in extrapolation.

It has been found that the number of periods in a signal is related to the training Time batch size and the accuracy of prediction. Fig. 7 show the percentage accuracy of all the signals with all the 8 training Time batches. Most of the pure as well as the mixed signals which are a combination of 2 signals, show accuracy higher than 50 percent after 2 Time batches, that is, after Time batch 0-24 and Time batch 0-48.

Signal S5 and Signal S6 are showing 100 percent accuracy after 2 and 1 time- batches respectively as their periodicity is small as compared to the other signals. In Signals S5 and S6, the patterns are repeating after 5 hours and 2 hours respectively. Although this is not the case for Signal S123 which is a mixed-signal of 3 signals. The accuracy is more than 50 percent after 5 time-batches.

Thus, it is clear that prediction accuracy is affected by the number of the periods and length of the period in the training data as well as the complexity of periodic patterns, while low error rates in periodic data can be overcome.

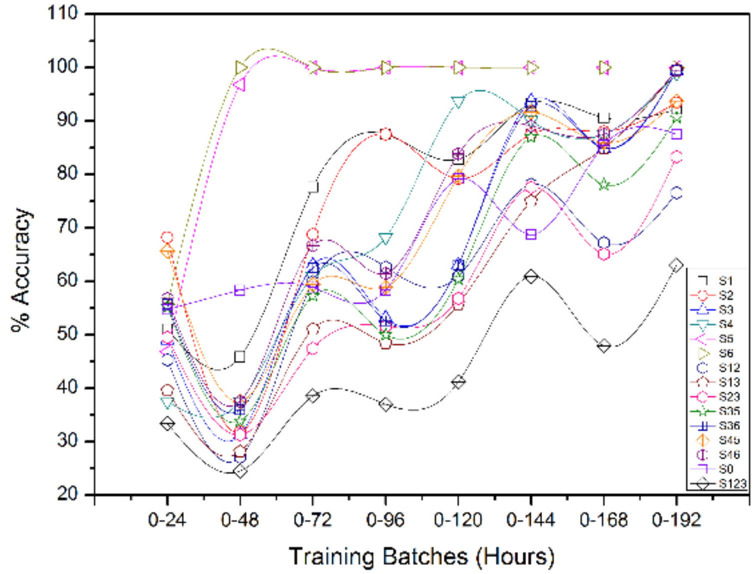

Fig. 7. Accuracy for all the signals with different Time batches

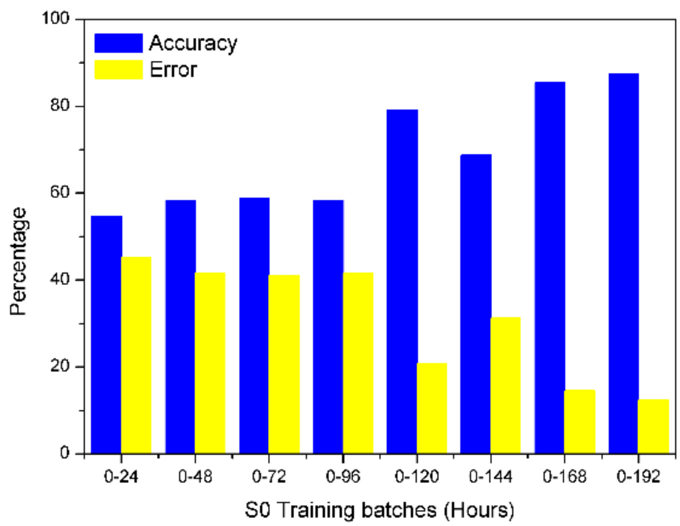

Fig. 8. Percentage accuracy and percentage error in extrapolation with different Time batches for Signal S0

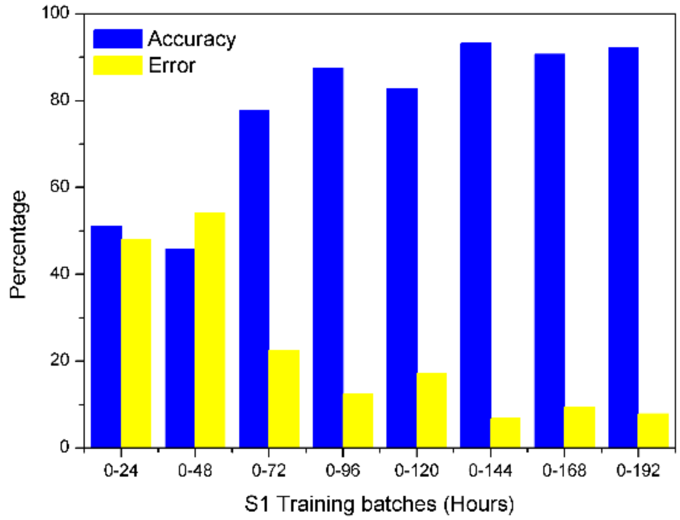

Fig. 9. Percentage accuracy and percentage error in extrapolation with different Time batches for Signal S1

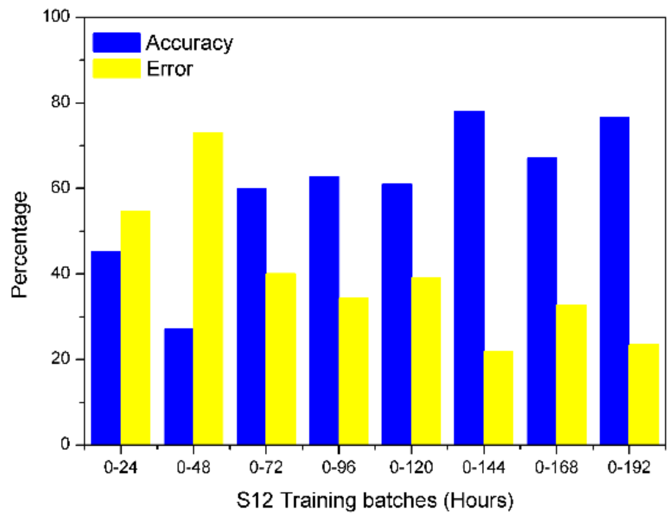

Fig. 10. Percentage accuracy and percentage error in extrapolation with different Time batches for Signal S12

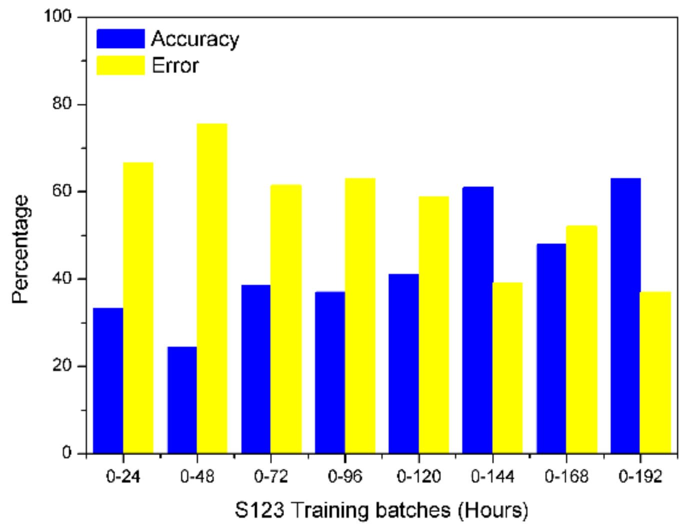

Fig. 11. Percentage accuracy and percentage error in extrapolation with different Time batches for Signal S123

5 Conclusion and Future work

This study sought to determine periodic patterns of behaviour in Web queries using signal processing. A typical user may exhibit complex periodic behaviour which, if understood, could be used to personalize and provide more relevant results to the user.

FFT was proposed to determine the strength and nature of periodic patterns for a user who may have such behavioural patterns. This can then be used as a generator to reproduce the periodic behaviour pattern in future. For example, a user who searches for sport every Saturday can have results personalized for sport on Saturdays.

The aim of this study was to determine how well this FFT approach can detect different kinds of patterns, using carefully controlled data, and the results indicate that it works reasonably well. The amount and variability of data clearly affect the outcome, but reasonably high prediction accuracies can be obtained with only the equivalent of a single week's data.

Other time forecasting methods (from machine learning) will be contrasted with this technique in future, and the approach will be tested with real-world data. Ultimately, this pattern detection will be used as part of a broader goal to personalise the search results for users.

References

1. Nagar, J., Suleman, H.: Investigating Per-user Time Sensitivity of Search Topics. In: TDDL/MDQual/Futurity@ TPDL (2017).
2. Horscroft, G., Nagar, J., Suleman, H.: Re-ranking the search results for users with time-periodic intents. In: Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists. pp. 239–247 (2018).
3. Song, Y., Guo, Q.: Query-Less: Predicting Task Repetition for NextGen Proactive Search and Recommendation Engines. In: Proceedings of the 25th International Conference on World Wide Web. pp. 543–553. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2016). https://doi.org/10.1145/2872427.2883020.
4. Mansouri, B., Zahedi, M.S., Rahgozar, M., Oroumchian, F., Campos, R.: Learning Temporal Ambiguity in Web Search Queries. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 2191–2194. ACM, New York, NY, USA (2017). https://doi.org/10.1145/3132847.3133129.
5. Odijk, D., White, R.W., Hassan Awadallah, A., Dumais, S.T.: Struggling and Success in Web Search. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 1551–1560. ACM, New York, NY, USA (2015). https://doi.org/10.1145/2806416.2806488.
6. Borisov, A., Markov, I., de Rijke, M., Serdyukov, P.: A Neural Click Model for Web Search. In: Proceedings of the 25th International Conference on World Wide Web. pp. 531–541. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2016). https://doi.org/10.1145/2872427.2883033.
7. Chuklin, A., Markov, I., Rijke, M. de: Click models for web search. Synth. Lect. Inf. Concepts, Retrieval, Serv. 7, 1–115 (2015).
8. Kliman-Silver, C., Hannak, A., Lazer, D., Wilson, C., Mislove, A.: Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In: Proceedings of the 2015 Internet Measurement Conference. pp. 121–127. ACM, New York, NY, USA (2015). https://doi.org/10.1145/2815675.2815714.
9. Liu, Y., Xie, X., Wang, C., Nie, J.-Y., Zhang, M., Ma, S.: Time-aware click model. ACM Trans. Inf. Syst. 35, 16 (2017).
10. Yi, X., Hong, L., Zhong, E., Liu, N.N., Rajan, S.: Beyond Clicks: Dwell Time for Personalization. In: Proceedings of the 8th ACM Conference on Recommender Systems. pp. 113–120. ACM, New York, NY, USA (2014). https://doi.org/10.1145/2645710.2645724.
11. Halvey, M., Keane, M.T., Smyth, B.: Time Based Segmentation of Log Data for User Navigation Prediction in Personalization. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence. pp. 636–640. IEEE Computer Society, Washington, DC, USA (2005). https://doi.org/10.1109/WI.2005.147.
12. Metzler, D., Jones, R., Peng, F., Zhang, R.: Improving search relevance for implicitly temporal queries. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 700–701 (2009).

13. Vu, T.T., Willis, A., Song, D.: Modelling time-aware search tasks for search personalisation. In: Proceedings of the 24th International Conference on World Wide Web. pp. 131–132 (2015).
14. Chowdhury, P., Srivastava, A.K., Dwivedi, B.N., Sych, R., Moon, Y.-J.: Study of multi-periodic coronal pulsations during an X-class solar flare. Adv. Sp. Res. 56, 2769–2778 (2015).
15. Li, D., Zhang, Q.M.: Quasi-periodic pulsations with multiple periods in hard X-ray emission. Mon. Not. R. Astron. Soc. Lett. 471, L6--L10 (2017).
16. Dillon, M.E., Woods, H.A., Wang, G., Fey, S.B., Vasseur, D.A., Telemeco, R.S., Marshall, K., Pincebourde, S.: Life in the frequency domain: the biological impacts of changes in climate variability at multiple time scales. Integr. Comp. Biol. 56, 14–30 (2016).
17. Samie, F., Paul, S., Bauer, L., Henkel, J.: Highly efficient and accurate seizure prediction on constrained iot devices. In: 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE). pp. 955–960 (2018).
18. Lafta, R., Zhang, J., Tao, X., Li, Y., Abbas, W., Luo, Y., Chen, F., Tseng, V.S.: A fast fourier transform-coupled machine learning-based ensemble model for disease risk prediction using a real-life dataset. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 654–670 (2017).
19. Lin, S., Liu, N., Nazemi, M., Li, H., Ding, C., Wang, Y., Pedram, M.: Fft-based deep learning deployment in embedded systems. In: 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE). pp. 1045–1050 (2018).
20. Seto, S., Zhang, W., Zhou, Y.: Multivariate time series classification using dynamic time warping template selection for human activity recognition. In: 2015 IEEE Symposium Series on Computational Intelligence. pp. 1399–1406 (2015).
21. Drutsa, A., Gusev, G., Serdyukov, P.: Periodicity in user engagement with a search engine and its application to online controlled experiments. ACM Trans. Web. 11, 9 (2017).