Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.070516
ISSN 2371-8838 Vol. 7 Num. 5

Prediction and Classification Model Based on Wordle's Date

Huang Yitai[1], Zhong Zeheng[2], Fang Zhaoyang[3]

[1]*School of Automotive Engineering, Wuhan University of Technology, Wuhan, China*
[2]*School of Computer and Artificial Intelligence, Wuhan University of Technology, Wuhan, China*
[3]*School of Science, Wuhan University of Technology, Wuhan, China*

Keywords: Wordle, Prophet, Apriori, Spectral Clustering

Abstract: Wordle is an interesting puzzle that gained huge popularity in early 2022. Studying the game's play data is important for its development and promotion. That's why our team has conducted research in this area. Firstly, we used Prophet Model to explain the reasons for the change in the number of reports and to predict the interval of change in the number of reports on a certain day in the future. Then, based on the word attributes that were tested by the Apriori Model, we use Spectral Clustering Model to classify the word by difficulty. With the above model, we obtained the results with good reliability and interpretability.

1. Introduction

Wordle, an interesting online fuzzy, achieves huge popularity in recent years. The target of the wordle is to guess a hidden five words in less than six times. Players receive feedbacks from yellow, green, or gray tiles. Different from the regular mode, Hard Mode also attracts many players. There are many reporting scores about daily results Twitter. The data set also contains the percentage of scores played in Hard Mode. Hence, it is of great significance to make full use of the data and find out the characteristic of each word as well ast he best strategy, which can get a better puzzle experience.[1]

2. Player Number Prediction Model

The number of players is an important indicator of how fun and promising a game is. Therefore, it is meaningful to study the number of reports.

We find that the number of reported results varies from day to day. To be able to explain this variation and to predict the number of reports for a future day, we develop the Prophet Model. We conclude that the data is cyclical, e.g., the number of reports is less on weekends than on weekdays. The Prophet Model is a good representation of the holiday effect and periodicity of the data with good interpretability.

2.1 Prophet Model

We employ the Prophet model to forecast and explain the trend of reported results. The Prophet is a modular regression model with interpretable parameters. Based on an additive model, Prophet are fit with yearly, weekly and daily seasonality. It can also eliminate the issue of delayed feedback. The fundamental mathematical relationship is formulated by:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \tag{1}$$

Where g(t) denotes the trend function which models non-periodic changes in the value of the time series; s(t) denotes periodic changes and h(t) denotes the holiday effects. Growth trend is the key component of the Prophet and plays a decisive role in the accuracy of the results. We utilize saturating growth to apply the logistic regression:

$$g(t) = \frac{C}{1 + \exp(-k(t-m))} \tag{2}$$

Where C denotes the carrying capacity that limits the maximum. k is the is the growth ratewhile m is the offset. In the real world problem, the carrying capacity C and the growth rate kare not always constant in the real world problem. Therefore, we replace fixed capacity withtime-varying parameter C(t) to improve practicality.

2.2 Consider Holiday Effects and Periodicity

The modelling of holidays or other events have significant impacts on the time series forecasting but do not follow a periodic pattern. Hence, it is necessary to include them as influencing factors in the forecast. For each holiday i, we denote Di as the set of past and future dates for that holiday. The parameter ki is assigned which is the corresponding change in the forecast as shown in the following equations:

$$Z(t) = [1(t \in D_1), ..., 1(t \in D_L)] \tag{3}$$

While taking

$$h(t) = Z(t)k \tag{4}$$

Additionally, the data set the dataset reveals a weekly periodicity. To better capture the periodic effects, Fourier series is introduced to the Prophet model. The weekly effects can be represented as follows:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right) \tag{5}$$

Where *P* denotes the regular period, and t denotes the time step.

2.3 Result and Analysis

Through Python, we get the future trend of reported results as illustrated in Figure 1. The confidence level is 95%.

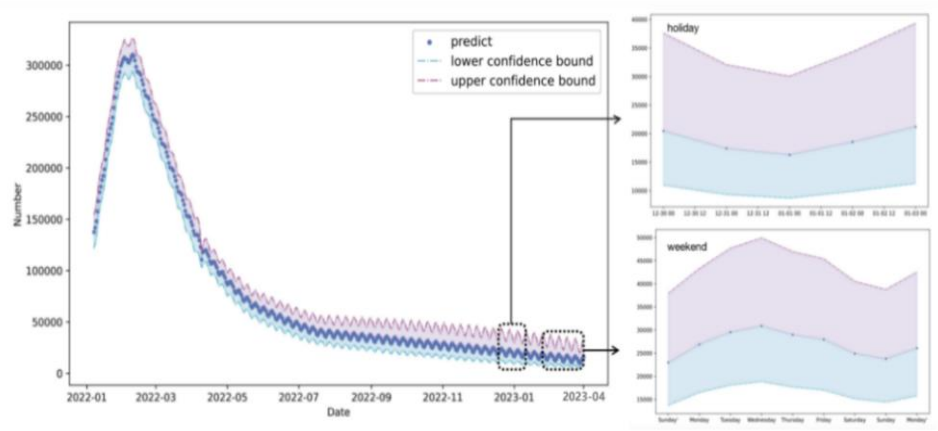

Figure 1: Prediction of Reported Results

From the figure, we find that the number of reported results is declining gradually and the rate of decline is also decreasing slowly. We can infer that people's interest in Wordle declines with the passage of time and the rise of other games. In April 1, 2023, the prediction interval is between 17505 and 19283 with a confidence of 95%.

On specific holidays, our result is in line with our expectations. For instance, the figure above shows that the number of reported results on New Year's Day, which is less than a few days before and after. Our results also reveals a weekly periodicity that is consistent with the original data.

3. Classify Words by Difficulty

3.1 Attribute selection

The difficulty of guessing varies from word to word and the attributes of a word count for a lot. We select 6 attributes to present the characteristics of a word:[2]

Frequency, Orthographic neighbors, Vowels, Root, Class, Letter frequency.

In order to verify whether the attributes we choose are related to the difficulty of the words. We take the percentage of the number of hard modes as an indication of the difficulty of a word. It is worth noting that the percentage of the number of difficult modes on a given day is affected by the difficulty of words on the previous day, that is, the data has a lag. We notice this when we were working with the data. On this basis, we select Apriori model to verify the relationship.[3]

The six attributes of words we selected, together with the percentage data form the set I = i1,i2, ...,i7, and T denotes a certain frequent item set, which consists of k items in I, and is denoted as T =t1,t2, ..., tk. The association rule is expressed as:

$$(X \in T) \Rightarrow (Y \in T) \tag{6}$$

It indicates that if there is an X item in the data relationship, there will be a Y item.

The above rule is bounded by Support (S) and Confidence (C), and the formula for calculating both are defined as follows:

$$\begin{cases} S(X \Rightarrow Y) = P(A \cup B) \\ C(X \Rightarrow Y) = P(B|A) = \dfrac{S(A \cup B)}{S(A)} \end{cases} \tag{7}$$

Where *S* denotes the frequency of rule occurrence whose value is determined by the ratio of the number of simultaneous occurrences. The confidence *C* indicates how real the rules are discovered. The higher the support *S* and the confidence *C*, the more intimate the relationship between this attribute and percentage. The reliability and usability of the rules can be evaluated and filtered by combining the set minimum support and minimum confidence. If it is higher than the set minimum confidence level, it is retained, otherwise it is discarded. The remaining retained items, which generate strong association rules, also satisfy the predefined minimum confidence threshold, thus mining the strong association rules.[4]

3.2 Spectral Clustering Model

To start with, the spectral clustering model uses Laplace matrices to reduce the dimensionality of the attributes. Then, the clustering transformed into a partitioning problem of spatially entitled undirected graphs. After partitioning the entitled undirected graph, the segmentation target becomes the high or low weight value, i.e., the weight value is high within the same class and low between different classes.[5]

For n samples $x_1, x_2, \ldots, x_n$ whose value is $x_{n,m}(1 \leq m \leq 6)$ for each attribute, considering the small amount of attached data, we choose the number of clusters as 4, i.e. k=4. [6] The specific implementation process is as follows:

1) Construct the weight matrix *G*: By using the K-nearest neighbor graph, we calculate the Chi-square distance and determine the value of *k*.

2) Construct the similarity matrix *S*: The similarity matrix is composed of the sum of the rows that in the weight matrix.

3) Calculate the normalized Laplace matrix: We define the Laplace matrix as: $L = S - G$. Then the normalized Laplace matrix is:

$$L' = S^{-\frac{1}{2}} L S^{-\frac{1}{2}} = 1 - S^{-\frac{1}{2}} G S^{-\frac{1}{2}} \tag{8}$$

Moreover, the objective of normalized spectral clustering is defined as:

$$\min_{F^T F = 1} Z = Tr(F') \tag{9}$$

Where *Tr* denotes the trace of the matrix and *F* is the class indicator matrix of the data set. Since *F* is a matrix composed of discrete values, it is necessary to use the eigenvector matrix *Q* composed of the eigenvectors, which corresponds to the smallest eigenvalues of L' as the relaxed continuous solution of the class indicator matrix *F*.

4) Calculate the feature vector: We calculate the feature vector corresponding to the maximum eigenvalue of k of the matrix L'.

5) Undirected graph segmentation: Each row of the feature vector matrix *Q* represents a sample. Then, we carry out K-means clustering algorithm on all samples to obtain clusters.

3.3 Result and Analysis

3.3.1 Result of Apriori

The Apriori model generates association rules between the frequent item sets. These rules indicate the likelihood of one item set appearing with another item set. The strength of each rule is measured by its confidence and support that have positive correlation with relationship as mentioned in section 3.1. The confidence and support of each index are calculated as shown in the following table.

Table 1: Results of the Apriori Model

Rule	Support (S)	Confidence (C)	Lift
Vowels	0.1583	0.6333	1.0556
Roots	0.0972	0.3846	1.1077
Class	0.0694	0.2809	1.2484
Frequency	0.0639	0.2584	1.0223
Letter frequency	0.0611	0.2418	1.1452
Orthographic neighbours	0.0500	0.2000	1.0746

From the Table 1, we find that the support and confidence are up to 0.1583 and 0.6333 respectively, indicating that vowels has a strong influence on the percentage of Hard Mode. The second influential attribute is the root. That is because the number of vowels exert impacts on the pronunciation and spelling. As for the orthographic neighbor, although it increases the difficulty of distinguishing words, it does not have significant influence on percentage of Hard Mode. The results of our model are correspond closely with common sense and experience.

3.3.2 Result of Spectral Clustering

The spectral clustering after Laplace transform where six attributes are downscaled into three main attributes. By analyzing the distribution of words, we find the clear classification characteristics, which will be helpful for the latter clustering.

We divide words difficulty into four levels: I-IV where level I refers the most difficult words as well as level IV refers the easiest words. We list some classification results in Table 2. We compare our results with www.twinword.com and Flesch–Kincaid Grade Level Formula (a technique to judge the readability and difficult level of a text). Our results reveals basically consistent with these references, indicating its high accuracy.

Table 2: Classification Results

Categories	Words
Level IV	catch, judge, night, enjoy, label, fewer, mummy. . . . .
Level III	carry, extra, usual, creak, needy, shame, yield . . . . .
Level II	inept, cynic, torso, alpha, amber, wrung, twine. . . . .
Level I	atoll, glyph, fungi, hinge, tiara, nymph, abbey. . . . .

According to our spectral clustering model, we select six attributes to classify the difficulty level. The central values in clustering are: Level I [3, 0, 2, 3, 4, 1] ; Level II [3, 1, 2, 3, 5, 2] ; Level III [4, 1, 1, 2, 5, 3]; Level IV [4, 1, 1, 1, 8, 4]. Level I represents the characteristics of low frequency, no root and large number of vowels while level II has the characteristics of middle frequency, root, and a bit of vowels. Level IV represents supper frequency, root, few vowels and many adjective words while level III has the characteristics of high frequency, root, and a few vowels. We also find that the frequency in Level I is significantly higher than other three levels. The number of vowels in Level IV is significantly lower than others, which shows the interpretability and reliability of our model.

4. Summary

Our model has certain advantages:

1) High data utilization: The model deeply explores the correlation between data and obtains good results.

2) Highly flexible: Thanks to the choice of our methodology, such as the Prophet, Apriori, the parameters in our models are easy to adjust, combined with strong parameter interpretation.

3) Widely adaptable: Our model can be easily generalized to the analysis of words in different length or other occasions, which shows its widely adaptable.

But there is still room for improvement:

1) Error effect: We make some simplifications in the model to facilitate the calculation, which may increase the error.

2) Missing other potentially relevant factors: When we select the word attributes, we may do not take all the relevant attributes into account.

References

[1] Anderson, Benton J., and Jesse G. Meyer. "Finding the optimal human strategy for wordle using maximum correct letter probabilities and reinforcement learning." arXiv preprintarXiv: 2202.00557 (2022).

[2] Stewart, Jeffrey, et al. "The relationship between word difficulty and frequency: A response to Hashimoto (2021)." Language Assessment Quarterly 19.1 (2022): 90-101.

[3] Kim Kwang Hyeon, Byung-Jou Lee, and Hae-Won Koo. "Analysis of the Risk Factors for De Novo Subdural Hygroma in Patients with Traumatic Brain Injury Using Predictive Modeling and Association Rule Mining." Applied Sciences 13.3 (2023): 1243

[4] Cong Yi. "Research on data association rules mining method based on improved apriori algorithm." 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE). IEEE, 2020: 373-376.

[5] Lei Jing, and Alessandro Rinaldo. "Consistency of spectral clustering in stochastic block models." (2015): 215-237.

[6] Bianchi Filippo Maria, Daniele Grattarola, and Cesare Alippi. "Spectral clustering with graph neural networks for graph pooling." International conference on machine learning. PMLR, 2020: 874-883.