Research on the Changes of Online Shopping Reviews based on Time Series Analysis

Xiaoyi Xu, Yue Jiang

School of Insurance, Central University of Finance and Economics, Beijing, 100081

Keywords: Online Shopping Reviews; Time Series Analysis; Detrended Cross-Correlation Analysis; Rescaled Range Analysis

Abstract: As e-shopping becomes one of the most successful innovation in the digital era, the online reputation of product occupies a more important role in the company's mind. In this paper, we used the provided dataset from the Sunshine Company to discover patterns of comments in each of the three categories of product, helping the Company to enter online market with its brand-new inventions. To study comments based on time-varying patterns, we established the time series models using Rescaled Range Analysis method and Detrended Cross-Correlation Analysis method. Particularly, the methods we used overcome the limitations of ordinary time series models which cannot accurately analyze non-stationary time series. Based on fractal system, our time series model researched the auto-correlation, persistence and cross-correlation of the time series. The use of Hurst exponent and DCCA method could be the biggest innovation of this article.

1. Introduction

Nowadays, online shopping led by Amazon hit a new milestone. The total market share of non-store, or online U.S. retail sales was higher than general merchandise sales for the first time in history, according to a report from the Commerce Department. Considering about the influence of ratings and reviews on the online market, a recent survey by Podium suggests that 93 percent of consumers do admit online reviews do impacting their purchasing decisions, while nearly two-third of them are willing to pay an additional 15 percent for the same product with high reputation. Moreover, 3.3 is the minimum star rating of a business consumers would consider engaging with. Therefore, it is crucial for the Sunshine company to identify key patterns, relationships, measures, and parameters in past customer-supplied ratings and reviews associated with other competing products, in order to gain a success for the three new products.

2. Rescaled Range Analysis

2.1 Hurst exponent

When British hydrologist H.E.Hurst (1900-1978) studied the relationship between water flow and storage capacity of the Nile Reservoir, he found that the use of biased random walks (fractal Brownian motion) can better describe the long storage capacity of the reservoir. Hrust exponent is an exponent describing the fractal characteristics of time series, which is suitable for biased random walks. The value range of it is often between 0 and 1. When the Hurst exponent is greater than 0.5, it means that the time series holds persistence, and the current rise is more likely to bring future rises; when the Hurst exponent is less than 0.5, it means the time series holds anti-persistence, the trend of the future is more likely contra the past; when the Hurst exponent is equal to 0.5, the time series is completely a white noise sequence, which has the same characteristics as the Brownian motion. The Hurst exponent can also be used to measure the auto- correlation of a time series. Nowadays, the Hurst exponent has become one of the most important time series models in the capital market.

The R\S method (Rescaled Range Analysis) proposed by H.E.Hurst (1951) originally studied the non-randomness of time series. Its central idea is to work out the Hurst exponent of a time series, in order to analyze whether the sequence holds persistence.

2.2 Methodology

At first, for a (partial) time series $\{x_t\}$ of length N, $t = 1,2,3 \cdots N$, calculate the cumulative deviate series $X(t,T)$:

$$M_n = \frac{1}{n}\sum_{t=1}^{n} x_t \qquad t = 1,2,3 \cdots N$$

$$X(t,n) = \sum_{i=1}^{t}[x_i - M_n] \qquad t = 1,2,3 \cdots u$$

Then we compute the range R :

$$R_n = \max_{1 \le t \le n} X(t,n) - \min_{1 \le t \le n} X(t,n)$$

and compute the standard deviation S:

$$S_n = \left[\frac{1}{n}\sum_{t=1}^{n}(x_t - Mn)^2\right]^{1/2}$$

Calculate the rescaled range $Rn\backslash Sn$ over all the partial time series of length N. We are supposed to find the following relationship between $Rn\backslash Sn$ and n:

$$Rn\backslash Sn \sim (cn)^H$$

Where *c* is constant. Log on both sides:

$$ln(Rn\backslash Sn) \sim H\ ln\ c\ +\ \ H\ ln\ n$$

With the known $Rn\backslash Sn$ and n, we can use the least squares method to fit out H, which is the Hurst exponent of the sequence $\{xt\}$.

2.3 Hurst exponent analysis

Apply the method to the data processed, we can get the Hurst exponents of different time series as follows:

Table 1. Hurst exponent of the hair-dryer time series

Product-id	Star- rating	Review- exponent	$W1(x) - K$	$W2(x) - K$	Discriminant
B00005O0MZ	0.62	0.71	0.69	0.51	0.63
B0009XH6TG	0.57	0.59	0.56	0.50	0.55
B00132ZG3U	0.66	0.57	0.64	0.64	0.62
Whole market	0.67	0.58	0.69	0.56	0.60

Table 2. Hurst exponent of the microwave time series

Product-id	Star- rating	Review- exponent	$W1(x) - K$	$W2(x) - K$	Discriminant
B0052G14E8	0.58	0.52	0.66	0.58	0.71
B0055UBB4O	0.54	0.54	0.56	0.69	0.78
B0058CLNBU	0.49	0.50	0.66	0.58	0.53
Whole market	0.64	0.65	0.76	0.56	0.57

Table 3. Hurst exponent of the pacifier time series

Product-id	Star- rating	Review- exponent	$W1(x) - K$	$W2(x) - K$	Discriminant
246038397	0.61	0.45	0.71	0.55	0.65
392768822	0.64	0.61	0.55	0.56	0.54
572944212	0.53	0.64	0.60	0.58	0.49
Whole market	0.51	0.54	0.51	0.54	0.61

As shown in the table, the Hurst exponent of most of the time series in the selected data is greater than 0.5, indicating that it holds persistence, which means that higher cur- rent reputation are more likely to signal a rise in future star-rating and review-exponent. In addition, for those products or markets where the Hurst exponent of $W1(x) - K$ is greater than $W2(x) - K$, positive star-rating and review-exponent are more likely to determine the future trend of reputation than bad ones, and vice versa. The Hurst exponent of the pacifier market is significantly smaller than the other two markets, even close to 0.5, which means the random walk. The Hurst exponent we get also mean that each time series we measure holds the characteristics of auto-correlation.

3. Detrended Cross-Correlation Analysis

The capital market is considered as a complex system because of interactions in eco- nomic variables. In other words, the time series we take from the capital market, especially from the sales data under this circumstance, is always non-stationary. Ordinary analysis methods cannot accurately analyze the correlation of non-stationary time series, and therefore we require the fractal theory. To solve the problem, Peng et al(1993) proposed DFA (Detrended Fluctuation Analysis) method to measure the multifractal exponents describing a non-stationary time series. However, the DFA method, which can be only used to measure one time series, is not able to calculate the cross-correlation between two time series. Something good took place. Podobnik and Stanley develop the DFA that studies the auto-correlation of one non-stationary time series into the DCCA method (Detrended Cross-Correlation Analysis) that studies the long-range cross-correlation between two non-stationary time series. After a lot of statistical analysis, the accuracy and usability of the DCCA method are gradually improved. Our team make some improvements on it.

3.1 Methodology

Assume two time series $\{x(t)\}$ and $\{y(t)\}$ of the same length T. Divide each time series into $V = \lfloor T/S \rfloor$ non-overlapping subinterval of equal length S, where $\lfloor x \rfloor$ means the largest integer not greater than x. Respectively, $S \in [10, T/4]$, for a more accurate result. So, we get V subintervals with S pieces of information each. For every subinterval $v = 1, 2, 3 \cdots V$, separately construct the sequence of cumulative sums.

$$X_v(t) = \sum_{i=(v-1)S+1}^{vS} x(i) \qquad t = 1,2,3 \cdots S$$

$$Y_v(t) = \sum_{i=(v-1)S+1}^{vS} y(i) \qquad t = 1,2,3 \cdots S$$

Considering that S is often not an integer, we do the same for the inverse order of $\{x(t)\}$ and $\{y(t)\}$ to ensure complete use of information. After that, we get $2V$ subin- tervals in total.

In each subinterval, fit $Xv\,(t)$ or $Yv\,(t)$ with a polynomial in order to get the corresponding fitting function. It is known from mathematical theory that fitting using a polynomial function of order n can be used to remove the trend component of a polynomial of order $n - 1$ in a signal sequence, so the polynomial function can be of any order. This paper uses the least square method to perform a first-order fit on two contour sequences. Use the fitting function to construct the fitted sequence $\{X'(t)\}$and $\{Y'(t)\}$. For $v = 1, 2, 3 \cdots 2V$, the covariance of residuals is defined as a function:

$$F(S, v) = \frac{1}{S}\sum_{t=1}^{S} |[X_v(t) - X'_v(t)][Y_v(t) - Y'_v(t)]|$$

Calculate the overall detrended fluctuation function:

$$F(S) = [\frac{1}{2V}\sum_{v=1}^{2V} F(S,v)]^{\frac{1}{2}}$$

The DFA method can be seen as a special case of the DCCA method when $\{x(t)\}$ is equal to $\{y(t)\}$. We can investigate the power-law relationship for different values of segment length S:

$$F\ (S) \sim S^{h}$$

And then, we use the least squares method to fit out H, which called the generalized Hurst exponent, similarly meaning to the Hurst exponent drawn by the RS analysis but measure the persistence between two time series. The cross-correlation coefficient is defined as follow:

$$\rho_{DCCA} = \frac{F_{xy}^2(S)}{F_{xx}(S)F_{yy}(S)}$$

The subscript of $F\ (S)$ indicates the time series it analyzes using the DCCA method. B.Podobnik and others have proven that $-1 < \rho DCCA < 1$. $\rho DCCA$ is equal to 1 only if $\{x(t)\} = \{y(t)\}$, and $\rho DCCA$ is equal to -1 only if $\{x(t)\} = -\{y(t)\}$.

3.2 Cross-correlation analysis

$\rho DCCA$ reflects the cross-correlation between two non-stationary time series. The re- sult is shown in the figure 0.5, 0.6, 0.7 in the appendix. The first three images of each row reflect the attributes of the product, while the last image reflects the whole market. From top to bottom are the cross-correlation between the time series of star-rating and discriminant, review-exponent and discriminant, star-rating and review-exponent. We use S-D, R-D, and S-R to replace the cross-correlation numbers of these three combinations, respectively.

Visually, S-D and R-D are very high in all combinations, showing the strong cross- correlation. For the hair-dryer, the whole market's S-D distribution is in the range [0.76,0.92], while the R-D distribution is in the range [0.83,0.93]. But for a single product, its S-D varies from [0.87,0.97] while the R-D varies from [0.85,0.95]. We can conclude that, compared to the whole market, for a single hair-dryer product, star-rating can affect its reputation more. And star-rating can be more efficient to discriminate whether the reputation of a single product is better than others.

Similarly, for the microwave, star-rating and review-exponent are almost equally im- portant for discriminating the reputation of a single product. For the pacifier, review- exponent may be a few more important. In general, the cross-correlation of a single product must be significantly higher than the whole market, showing that a good product (with the largest quantity of comments) will have a high dependence on both star- rating and review-exponent.

Expectedly, the combination of star-rating and review-exponent shows very strong cross-correlation, expressed as extremely high S-R. Not surprisingly, a customer who is willing to score high will always write reviews that praise the product.

3.3 Persistence analysis

Similar to section 2.2, the formula 0.8 and 0.9 reflect the persistence exponent between two time series. But this time we do not draw the S-R graph because We have fully understood their correlation before.

We need to explain the figures. The figure is about the partial persistence exponent, which means that the two time-series do not correspond exactly in time, but a few days part. The horizontal axis is the number of days. When the legend shows 'S to D', it means that the star-rating is before the discriminant, for a single product or the whole market. We use this to study the continuous effect of star-rating and review-exponent on reputation (expressed ad discriminant) in time.

For the hair-dryer market, the persistence exponent rise within a certain time and then fall. The inflection point is about the tenth day. For a single product, it shows the opposite trend. It may mean that there is a long feedback period for hair-dryer.

Similarly, the other two markets have the same trend, but not for single products. For the microwave, the movement pattern of persistence exponent is not obvious, but for the pacifier, it falls sharply from the beginning, meaning that there is a short feedback period.

No matter who is before, the persistence exponent of 'S to D' and 'D to S' is similar. We can draw the conclusion that the effect of star-rating on reputation is similar to the effect of reputation on star-rating. Higher star-rating often inspire future customers to praise the product, and excellent reputation can attract higher star-rating. Virtuous circle! The same thing happens on the combination of review-exponent and discriminant as well

4. Conclusion

To study comments based on time-varying patterns, we established the time series models using Rescaled Range Analysis method and Detrended Cross-Correlation Analysis method. Particularly, the methods we used overcome the limitations of ordinary time series models which cannot accurately analyze non-stationary time series. Based on fractal system, our time series model researched the auto-correlation, persistence and cross-correlation of the time series. We can see that there would be Spillover Effect and Herd Behavior in the market. In the low star-rating comments, the negative descriptors will become more and more obvious, with other shortcomings bursting out, which means that costumers are more likely to write negative reviews for those low star- rating products. As the same, high star-rating comments can be a heavenly gift for the product, causing linear increase in the number of positive reviews.

References

[1] Hurst H E. The longterm storage capacity of reservoirs. Transactions of the Ameri- can Society of Civil Engineer 116, 1951

[2] C K Peng, S V Buldrev, correlation: implications for 47 (5): 3730-3733.A L Goud- berger, et al. Finite size effects on long range analyzing DNA sequences [J]. Physical Review E, 1993

[3] Podobnik B, Stanley HE. Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series. [J]. Physical review letters, 2008, 100(8): 084102

[4] Guangxi Cao, Yan Han, Yuemeng Che, Chunxia Yang, Multifractal detrended cross- correlation between the Chinese domestic and international gold markets based on DCCA and DMCA methods [J]. Modern Physics Letters, 2014

[5] A survey on opinion mining and sentiment analysis: Tasks, approaches and applications [J]. Knowledge-based Systems, 2015, 89: 14-46