Research Journal of Applied Sciences, Engineering and Technology 6(24): 4563-4568, 2013
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: December 15, 2012 Accepted: January 17, 2013 Published: December 25, 2013

The Research on Data Mining of Slim Life Mode Based on Cycle Behavior

[1]Liu Xianglin and [2]Yao Binbin
[1]Physical Education Department, Dalian Jiaotong University, Dalian 116028, China
[2]Physical Education Department, Anqing Teachers College, Anqing 246000, China

Abstract: In this study, data mining of slim life mode based on cycle behavior is propsed. The mining of the periodic behavior is divided into four stages. The first two stages is data pre-processing stage: Firstly, parsing stay point sequence from data sequence of the original location history. Here stay point represent the geographic area to a person's stay for some time; Secondly, cluster mining the sequence of stay point, find out the significant places, such as company, supermarket, home location, etc. Thirdly, mining periodic on the significant places. Take a place as a reference point; abstract the original location history data into binary sequence by the location point in or out the place. Then, combination two popular signal processing method fast Fourier and autocorrelation find the periods of every place. Fourthly, mining the periodic behavior of the places with the same periods, in this article, first construct the periodic behavior probabilistic model, then use the method based on the hierarchical clustering to mining the periodic behavior between different places. At last, an example is introduced.

Keywords: Cycle behavior, data mining, slim life mode

INTRODUCTION

Balanced diet cycle theoretical basis rather complex, once owned by the central research department spent three years as many as 30,000 when the track was summed individual clinical conclusions. In simple terms, can be understood: body fat at different times and at different stages have different degrees of change, if we can fully appreciate the changes in the most accurate time you can control fat increase or decrease, but actually does not do modern medicine to the exact hour of the fat changes to the law.

The normal sequence of $x(n)$, n = 0, 1…N-1 discrete Fourier transform is a complex sequence $X(f)$:

$$X(f_{k/N}) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi kn}{N}}, \quad k = 0,1...N-1$$

Where in, subscript k/N represent each factor produced by frequency. Therefore, in this paper, we will use $F(x)$ to represent the Fourier transform. For real signals, the Fu Liye factor is symmetric (more specifically, they are symmetric complex conjugate). Fu Liye converted to use the more complex sinusoidal function:

$$s_f(n) = \frac{ej2\pi fn / N}{\sqrt{N}}$$

are linearly combined to represent the original signal. Therefore, the Fu Liye factor in signal x projection to them, keep the amplitude and the phase of the sine function.

We can use the inverse Fourier transform from frequency domain to time domain $F^{-1}(x) \equiv x(n)$:

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(f_{k/N}) e^{\frac{j2\pi kn}{N}}, \quad n = 0,1...N-1$$

As shown in Fig. 1, in the inverse conversion process, we discarded some factor; the result will be the original sequence of approximate value. By selecting the need to record the factor, we can put the Fu Liye transformation is applied to many fields, such as information compression, removal of noise.

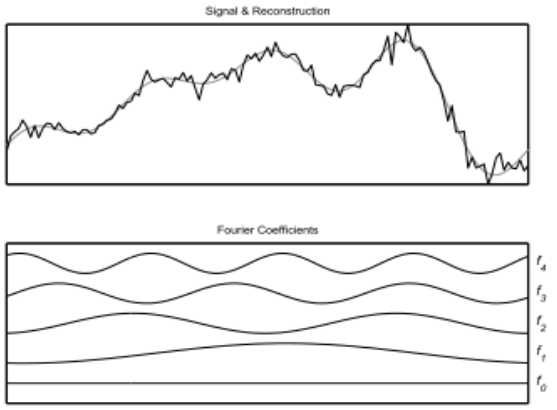

Fig. 1: Reconstruction signal from the original 5 Fu Liye factors

Corresponding Author: Yao Binbin, Physical Education Department, Anqing Teachers College, Anqing 246000, China

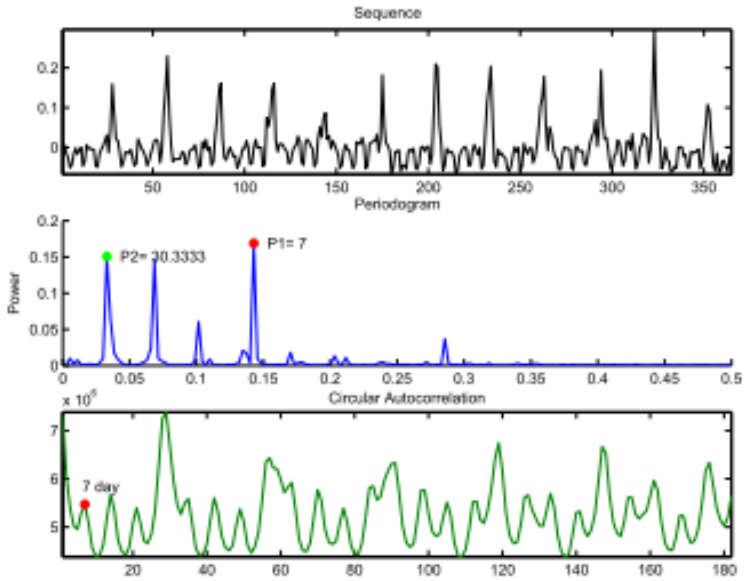

Fig. 2: Period gram and correlation diagram of the sequence

Power spectral density estimation: In order to detect the time sequence cycle, need to study its Power Spectral Density (PSD). Power spectrum of the signal at each frequency signal power expectations. The cycle is the frequency of inverse operation, by identifying with most energy in the frequency, we can find that most of the explicit periodic. Estimation of PSD commonly used in two ways: the cycle graph and cyclic autocorrelation. The two methods can be used to calculate the sequence of discrete Fu Liye transform (Lee *et al*., 2008a; Giannotti *et al*., 2007).

Cycle diagram: Suppose X is a sequence of Fu Liye transform. Cycle diagram P can use each of the Fu Liye factor to calculate the square length:

$$P(f_{k/N}) = \left\| X(f_{k/N}) \right\|^2 \quad k = 0,1...\left\lceil \frac{N-1}{2} \right\rceil$$

Which, according to the theorem, loss, because the data window effect, we can only detect the maximum signal frequency more than half the frequency (Xia *et al*., 2006; Jeung *et al*., 2008; Liao *et al*., 2005), in order to find the K dominant period, we need to pick out the cycle graph of the maximum K value.

Each cycle graph element expressed in the frequency of k/N or N/k power cycle. Rather, each discrete Fourier "box" corresponding to a series of cycles. That is to say, the corresponding cycle factor X($f_{k/N}$) ($\frac{N}{k}$ ... $\frac{N}{k-1}$). Can be found, along with the cycle length increased, cycle graph of the resolution is very low (Ye *et al*., 2009; Li *et al*., 2008; Cao *et al*., 2007).

With the increase of cycle, the cycle that reduced precision of reason is the Fourier "box" length increases. Another relevant cause of spectrum leakage, which prompted the length is not a discrete Fourier "box" length is an integer multiple of the frequency dispersion to the entire spectrum of. This will lead to a cycle of "false alarms". However, cycle diagram can still for important short cycle provides precise instructions. And, through the cycle graph, can easily through the detection of Fu Liye's statistical property to automatically extract important cycle (Lee *et al*., 2008b; Yan *et al*., 2003; Krumm and Horvitz, 2006; Giannotti *et al*., 2006).

Cyclic autocorrelation: Second kinds of estimation of time series X explicit cycle is the method of cyclic autocorrelation (ACF), which tests a series of varied time τ value similarity:

$$ACF(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} x(\tau)x(n+\tau)$$

Therefore, since the correlation in the form of a convolution in time domain, we can avoid the square root calculation and using frequency domain normal Fu Liye converted to compute its value:

$$ACF = F^{-1} < X, X^* >$$

The symbol * indicates complex conjugate: Cyclic autocorrelation vs cycle diagram, gives more fine-grained cycle detection, therefore, it can be more accurate detection of long period. However, due to the following reasons, it is not suitable for automatic periodic search (Elfeky *et al*., 2005; Bar-David *et al*., 2009; Li *et al*., 2010):

- Automatically discover important peak compared to cycle graph is difficult. At present based on autocorrelation method requires manual setting effective threshold.
- Even offers effective threshold, will still find plenty to meet the conditions of the period. Therefore, the need for additional operations to eliminate the "false alarm".
- The high frequency and low amplitude events could be compared to high amplitude event, is not important, although this rarely occurs. As shown in Fig. 2, 7 day cycle in the autocorrelation graphics with low amplitude was not important, however, in the cycle graph, on day 7 of the cycle is very clear (Zhao and Mao, 2011; Zhu, 2011).

From what has been discussed above, we can realize the fact that although the period gram and correlation could not separate with all information of the spectrum, however, through the combination of these two methods, you may find satisfying all of the spectrum information method. The following individuals in important places cycle detection, with a combination of Fu Liye transform and autocorrelation for cycle method.

- **Binary sequence periodic detection:** For individuals with an important place, we put forward a kind of the important place to find potential cycle method. To an important place as a reference point, moving sequences can be converted into a binary sequence: $B = b_1b_2\ldots b_n$, which, at the time stamp is i, individuals in the important sites, $b_i = 1$; otherwise, $b_i = 0$.

The above said, we used a combination of Fu Liye transform and autocorrelation for cycle method to find the binary sequence in the cycle.

In the Discrete Fourier Transform (DFT), sequences of $B = b_1b_2\ldots b_n$ converted to N complex sequence $X_1, \ldots, X_n$. For factor X_k, cycle graph is defined for each Fu Liye factor square length: $F_k = \|X_k\|^2$. Type, F_k is the frequency of the K power. In order to specify which frequency is effective, we need to set a threshold and labeled to exceed this threshold frequency.

Through the following method to determine the threshold. Let B' be a sequence of B a random permutation. Because the B' should not have any periodic, even the largest power in the sequence does not indicate periodicity, therefore, we remember its maximum power is P_{max} and only in the sequence of B is higher than that of P_{max} frequency corresponds to a real cycle. In order to make important frequency confidence rate reached 99%, we repeat the above random permutation sequence 100 times and record every permutation sequence maximum power. The 100 experiments in the ninety-ninth largest power values are used as the assessment of the power threshold.

For more than the power threshold F_k we still need to determine the exact period of time domain, because in the frequency domain of the value K corresponding to the time domain $[\frac{N}{k}, \frac{N}{k-1})$ between a series of cycle. In order to determine the cycle, we use cyclic autocorrelation, to assess a sequence in a different tag sequence value similarity: $R(\tau) = \sum_{i=1}^{n} b_\tau b_{i+\tau}$.

Therefore, the cycle graph are given in the period [l, r), we pass the data into a square functions to test in $\{R(l), R(l+1), \ldots, R(r-1)\}$ if there exists a peak value. If the function returns the result is, periodic region of a concave, heralded a peak exists, we return to the $t^* = \arg\max_{l \le t < r} R(t)$ as a probe of cycle. Therefore, we return to the $t^* = \arg\max_{l \le t < r} R(t)$ as a probe of cycle.

Periodic behavior mining of important locations: By data processing and periodic detection to obtain each important location in the cycle after cycle, we next discuss behavior mining. We will have the same cycle important sites focus to obtain more concise and valuable cycle behavior. However, due to a behavior there is only a part of the mobile; the same cycle may have a number of periodic behaviors. For example, in Windows Mobile, with two daily behavior. A place in the school, another place in summer. However, for a long time movement and a daily cycle, we don't really know what moving much cycle behavior and the number of days to a periodic behavior. However, we observed that with the same cycle behavior of "day" has the same spatiotemporal patterns. Therefore, you can use clustering method to mining period. Through the application of a model to measure two "days" of the distance between them, we can further packet days to several cluster and each cluster represents a periodic behavior. As in the small example, "school" should be grouped into a cluster; summer should be grouped into a cluster of. Therefore, periodic behavior of mining faces a major issue is to establish the model of periodic behavior, put forward the model based on clustering distance function.

Cycle behavior model: First of all, we picked out all the important places individuals with periodic T. Through the combination with the same cycle personal important locations, we can get different important sites and the period between behavior knowledge. For example, we can summarize Xiaoqiang everyday behavior "in company 9:00-18:00, 20:00-8:00 in the dormitory".

With period T personal important locations in the set $O_T = \{o_1, o_2, \ldots, o_d\}$, we use o_0 except for important locations outside the $o_1, o_2, \ldots, o_d$ position. For each position of $LOC = loc_1\ loc_2\ \ldots\ loc_n$ sequences, we generate a corresponding mobile symbol sequence S =

S_1, S_2, …, S_n when loc_i o_j, $s_i = j$. *S* is further divided into $m = \left\lfloor \frac{n}{T} \right\rfloor$ segments. We use I_j to express J, $t_k(1 \le k \le T)$ said a cycle within the first k relative time stamp. The $I_k^j = i$ object in paragraph j t_k in O_i, in O_i in. j For example, for a period of T = 24 hours, on behalf of " one day ", said t_9 day 9:00 and $I_9^5 = 2$ said object in fifth days at the o_2 9:00, naturally, we can use the spatial and temporal distribution to establish the probability model.

Maximum Likelihood Estimation (MLE): In statistics, Maximum Likelihood Estimation (MLE) is a statistical model parameter estimation method. When we are in a data set using a probabilistic model, maximum likelihood estimation for model parameter estimation. Usually, the data set and the underlying probability model, maximum likelihood method by generating a can make the observation data of maximum probability distribution to select the model parameter values (parameters to maximize the likelihood function).

If a contains the *n* observations of IID samples x_1, x_2, …, x_n, from an unknown probability distribution function $f_0(x)$, speculated that the function FO belongs to a called parametric model to determine the distribution of $\{fx|\theta), \theta \epsilon \Theta\}$, then f_0. The value of θ_0 is unknown and is regarded as the true value of parameter. So looking into some as close to the true value of the estimated value is feasible. Here observed variables x_i and θ are considered as vector parameters.

In order to use maximum likelihood estimation, first specify all observed values of joint distribution function:

$$f(x_1,...,x_n \mid \theta) = f(x_1 \mid \theta) \times f(x_2 \mid \theta),..., f(x_n \mid \theta)$$

Now, we get another view of this function, the observed variable x_1, x_2, …, x_n, as the function of the fixed parameter, θ is regarded as the function of the variable and allow free variation. From this point of view to see the distribution function is called a likelihood function:

$$L(\theta \mid x_1,...,x_n) = f(x_1,...,x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

In the actual application is commonly used in both sides of logarithm. Get the following formula:

$$\ln L(\theta \mid x_1,...,x_n) = \sum_{i=1}^{n} \ln f(x_i \mid \theta)$$

The $\hat{t} = \frac{1}{n} \ln L$, $\ln L(\theta|x_1, \ldots, x_n)$ called the log likelihood and $\hat{t}$ called the average log-likelihood. We then called maximum likelihood is the largest average log likelihood, i.e.:

$$\hat{\theta}_{mle} = \arg\max_{\theta \in \Theta} \hat{l}(\theta \mid x_1,..., x_n)$$

For many models, maximum likelihood estimation can be observed via an x_1, x_2, …, x_n, clear function and solving them; for the other models, there is no problem of maximizing the closed form solution, then the MLE through optimization method for solving.

Space-time distribution matrix: Let $T = \{t_1, t_2, \ldots, t_T\}$ be a cycle time stamp collection, x_k said in the time stamp t_k when choosing the reference point category of random variables. $P = [p_1, p_2, \ldots, p_T]$ is a space-time distribution matrix, wherein each column expressed as:

$$P_k = [p(x_k = 0), p(x_k = 1),..., p(x_k = d)]^T$$

independent category distribution vector, it satisfies the $\sum_{i=0}^{d} p(x_k = i) = 1$.

Now, assume that I^1, I^2, …, I^l has the same periodicity, segment set $I = U_{j=1}^{l} I^j$ probabilities can be obtained by some distribution matrix P to generate:

$$P(I \mid P) = \prod_{I^j \in I} \prod_{k=1}^{T} p(x_k = I_k^j)$$

According to the Maximum Likelihood Estimation (MLE), the optimal model can be defined as the following in maximum likelihood problem optimal solution:

$$\max_P \{L(P \mid I) = \ln P(I \mid P) = \sum_{I^j \in I} \sum_{k=1}^{T} p(x_k = I_k^j)\}$$

Formula solution:

$$p(x_k = i) = \frac{\sum_{I^j \in I} 1_{I_k^j = i}}{|I|}$$

The 1_A representation and event A associated with the indicator function. That is to say, $p(x_k = i)$ is a reference point in time relative to the o_i t_k I frequently.

Periodic behavior: I represent a section of the collection; all the I segment of the periodic behavior is expressed as $H(I)<T, P>$. T said cycle length; P is from the equation learning the space-time distribution matrix. We further make the |I| of said cover this cycle behavior of all the number of segments.

Cycle behavior mining: Periodic behavior clustering distance function. By periodic behavior is defined, we can segment set on the estimation of periodic behavior, now a section of the set $\{I^1, I^2, \ldots, I^m\}$, we need to find those segments from the same period. Assuming there is k a potential periodic behavior, each one found in

parts of the move, all of the paragraph should be divided into k groups and each group corresponds to a periodic behavior.

To solve the problem of the potential method is to use a clustering method. In order to use this method, two cycle behavior of distance metric needs to be defined. As a behavior is expressed as a <*T*, *P*> and T is fixed, so the distance from their space-time distribution matrix to determine the. Further, the two cycle behavior between small distance indicates that contains each behavior of some may be produced from the same period.

There are many methods to measure the temporal and spatial distribution of matrix P and Q of the distance between them. Here, we assume that different time stamp on the variables are independent, we propose the use of the famous Kullback-Leibler differences as distance measurement:

$$\mathrm{KL}(P\|Q)=\sum_{k=1}^{T}\sum_{i=0}^{d}p(x_k=i)\ln\frac{p(x_k=i)}{q(x_k=i)}$$

When $KL(P||Q)$ is very small, which means that the P and Q distribution matrix is similar to, or are different.

Notice that when $p(x_k = i)$ or $q(x_k = i)$ has a probability of 0, the $KL(P||Q)$ value is infinite. In order to avoid the occurrence of such a situation, we all reference points $p(x_k = i)$ and $q(x_k = i)$ increase a background variable u:

$$p(x_k=i)=(1-\lambda)p(x_k=i)+\lambda u$$

where, λ is a small parameter 0<λ<1.

In order to further from a statistical point of view the above method can solve our problem, we returned to our proposed model. Because I am a distribution matrix generated by the P segment set and then $KL(P||Q)$ can be further developed:

$$\begin{aligned}KL(P\|Q)&=\sum_{k=1}^{T}\sum_{i=0}^{d}p(x_k=i)\ln p(x_k=i)\\&\quad-\sum_{k=1}^{T}\sum_{i=0}^{d}p(x_k=i)\ln q(x_k=i)\\&=-H(P)-\sum_{k=1}^{T}\sum_{i=0}^{d}\frac{\sum_{I^j\in I}1_{I_k^j=i}}{|I|}\ln q(x_k=i)\\&=-H(P)-\frac{1}{|I|}\sum_{I^j\in I}\sum_{k=1}^{T}\ln q(x_k=I_k^j)\\&=-H(P)-\frac{1}{|I|}\ln P(I|Q)\end{aligned}$$

where, $H(P)$ is the P entropy, can be seen as a constant. Therefore, Kullback-Leibler difference measurement segment set I can be generated by Q possibility distribution matrix. In our clustering algorithm, for Q to make a choice, we simply choose the maximum likelihood $P(I|Q)$ the Q.

Now, suppose we have two periodic behavior, $H1$ = <*T*, *P*> and $H2$ = <*T*, *P*> We define two acts as the distance between:

$$dist(H_1,H_2)=KL(P\|Q)$$

Periodic behavior clustering algorithm: Assuming the existence of K potential periodic behavior, there are many ways to packet segment set to K cluster. However, potential cycle number is usually not known. We propose a hierarchical clustering method to packet segment set and determine the optimal number of periodic behavior. In hierarchical clustering in each iteration, with a minimum distance of two clustering fusion. We use a representation error to monitor the quality of clustering. When the cluster number from K to k-1, if the representation error suddenly increases, suggesting that K may be the correct period number.

CONCLUSION

Of course, the equilibrium cycle diet is just an application of the theory; with the deepening of the late, fat change will also make other applications of various forms. Ideally, hierarchical clustering process, from the same behavior section first fusion, because they have the smallest distance. Therefore, we use clustering in the section is in a special time are concentrated into a separate individual important sites to determine the clustering is good. Therefore, a natural representation of error measurement method is to estimate the quality of clustering.

REFERENCES

Bar-David, S., I. Bar-David, P.C. Cross, S.J. Ryan and W.M. Getz, 2009. Methods for assessing movement path recursion with application to African buffalo in South Africa. Ecology, 90: 2467-2479.

Cao, H., N. Mamoulis and D.W. Cheung, 2007. Discovery of periodic patterns in spatio- temporal sequences. IEEE T. Knowl. Data En., 19(4): 453-467.

Elfeky, M.G., W.G. Aref and A.K. Elmagarmid, 2005. Warp: Time warping for periodicity detection. Proceeding of 5th IEEE International Conference on Data Mining (ICDM).

Giannotti, F., M. Nanni and D. Pedreschi, 2006. Efficient mining of sequences with temporal annotations. Proceeding of SIAM Conference on Data Mining, pp: 346-357.

Giannotti, F., M. Nanni, F. Pinelli and D. Pedreschi, 2007. Trajectory pattern mining. Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp: 330-339.

Jeung, H., Q. Liu, H.T. Shen and X. Zhou, 2008. A hybrid prediction model for moving objects. Proceeding of the IEEE 24th International Conference on Data Engineering (ICDE 2008). Cancun, pp: 70-79.

Krumm, J. and E. Horvitz, 2006. Predestination: Inferring destinations from partial trajectories. Proceeding of the 8th International Conference on Ubiquitous Computing (UbiComp 2006). Orange County, CA, USA, September 17-21, pp: 243-260.

Lee, J.G., J. Han, X. Li and H. Gonzalez, 2008a. Traclass: Trajectory classification using hierarchical region-based and trajectory-based clustering. Proc. VLDB, 1(1): 1081-1094.

Lee, J.G., J. Han and X. Li, 2008b. Trajectory outlier detection: A partition-and- detect framework. Proceeding of IEEE 24th International Conference on Data Engineering (ICDE), pp: 140-149.

Li, Q., Y. Zheng, X. Xie, Y. Chen, W. Liu and W.Y. Ma, 2008. Mining user similarity based on location history. Proceeding of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, New York.

Li, Z., B. Ding, J. Han, R. Kays and P. Nye, 2010. Mining periodic behaviors for moving objects. Proceeding of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, pp: 1099-1108.

Liao, L., D.J. Patterson, D. Fox and H. Kautz, 2005. Building personal maps from gps data. Proceedings of IJCAI Workshop on Modeling Others from Observation (IJCAIMOO 05), pp: 249-265.

Xia, Y., Y. Tu, M. Atallah and S. Prabhakar, 2006. Reducing data redundancy in location-based services. Proceeding of 2nd International Conference on Geosensor Networks, Boston, MA, pp: 30-35.

Yan, X., J. Han and R. Afshar, 2003. CloSpan: Mining closed sequential patterns in large datasets. Proceeding of SDM, pp: 166-177.

Ye, Y., Y. Zheng, Y. Chen, J. Feng and X. Xie, 2009. Mining individual life pattern based on location history. Proceeding of International Conference on Mobile Data Management: Systems, Services and Middleware, pp: 1-10.

Zhao, L. and Y.X. Mao, 2011. GOBO: A sub-ontology API for gene ontology. IEIT J. Adap. Dynam. Comp., 1(1): 35-40.

Zhu, X.D., 2011. Block correlations directed multi-copies data layout technology. IEIT J. Adap. Dynam. Comp., 2011(1): 33-38.