Research on Stereo Library Based on Homogeneous Markov Chain

Shunkai Sun, Jie Li, Qi Xu, Haihua Lu*, Haobo Cui, Xue Shen, Yingxiao Chi and Jingjie Liu
Ningbo Cigarette Factory, China Tobacco Zhejiang Industrial Co., Ltd., Ningbo 315504, China

*Corresponding author e-mail: luhaihua@zjtobacco.com

Abstract. The intelligent inventory has not yet been implemented in the automated three-dimensional library. The warehouse basically relies on the accounts of the warehouse. It is really necessary to check the pallets manually with the help of a stacker, or venture into the aisle to check on the spot, or install radio frequency identification (RFID) on the pallet. The radio frequency recognition device on the stacker's loading platform obtains the information of a specific pallet to achieve a physical inventory. After investigation, the results that can pass the three-dimensional library system automatic library are still blank. For the three-dimensional library, the intelligent inventory is of great significance to the actual production. It can automatically identify whether the goods in the designated cargo space are consistent with the storage according to the needs of the warehouse manager, so the demand for the intelligent inventory is very urgent.

1. Introduction
The three-dimensional warehouse has remarkable characteristics, especially the advantages of fast cargo storage and storage, large storage capacity, and high space utilization. It has become a key link in logistics and is widely used in various manufacturing enterprises. Mainly distributed in the fields of medicine, tobacco, food, manufacturing, coal industry, military department, airport and port. As the scale of automated three-dimensional warehouses has increased, deeper requirements have gradually emerged. An important function in the warehouse management system, such as "distribution warehouse", appears to be stretched in the automated three-dimensional warehouse system. How to confirm a certain point in time, the warehouse Whether the physical inventory is consistent with the information, and whether the goods in the warehouse are checked against the information in the database, has become a very important and realistic issue.

2. Homogeneous Markov chain definition
Given a Markov chain \( \{X_n, n=0, 1, 2, \ldots\} \), if its one-step transition probability \( p_{ij}(m) \) has nothing to do with the time starting point \( m \),

\[
p_{ij}(m) = P\{X_{m+1} = i | X_m = j\} = p_{ij}
\]

(1)

Then call \( \{X_n, n=0, 1, 2, \ldots\} \) as a homogeneous Markov chain.
Stationary distribution of a finite Markov chain: a vector

\[ \pi = (\pi_0, \pi_1, \cdots, \pi_{k-1})^T \]  

(2)

It is called the stationary distribution of a finite Markov chain, if it satisfies the condition:

\[ \pi \geq 0, \sum_{j=0}^{k-1} \pi_j = 1 \]  

(3)

\[ P\pi = \pi, \text{That is} \sum_{j=0}^{k-1} p_{ij} \pi_j \]  

(4)

The existence of a stationary distribution vector: There must be at least one stationary vector for an irreducible and non-periodic Markov chain with k states.

For any irreducible and aperiodic Markov chain with k states, if an initial probability vector \( X_0 \) is arbitrarily given.

\[ \lim_{n \to \infty} \lim_{n \to \infty} X_n P = \pi \]  

(5)

Here is the stationary distribution corresponding to the probability transition matrix \( P \), and the above formula is to limit a certain vector norm. The above formula provides a theoretical basis for us to find the stable distribution of the Markov chain. The most important thing to apply the Markov chain is to find the stable distribution of the Markov chain.

3. Virtual Inventory Design

This article's warehouse management system mainly includes three functional modules: data preprocessing module, data learning module and abnormal operation module.

3.1. Data pre-processing

The data preprocessing module mainly includes the data acquisition submodule and the preprocessing submodule.

1) Data acquisition sub-module. In this article, the warehouse management system uses SQL statements as the data source for abnormal operations. Therefore, it is necessary to collect the high-shelf database SQL operation requests in the database log files.

2) Preprocessing sub-module. In this module, the collected data is mainly processed and processed to make it conform to the input of the next module. According to the requirements of the PostgreSQL database, all collected data needs to be processed into a sequence. Therefore, the keywords in the SQL request need to be extracted and combined into a learning sequence. In the implementation process, since the keywords of the SQL statement are usually English words with multiple letters, the operation is more complicated in the learning and operation process. In order to optimize the execution efficiency of the algorithm, the keywords are statistically classified and replaced by a single English letter.

3.2. Data learning

The data learning module includes a sequence library establishment sub-module and a POSTGRESQL-based behavior feature extraction sub-module.

1) Sequence library building module. For the preprocessed sequence, define \( W \) short sequences of different lengths, and set the corresponding frequency weights.

\[ y = (y_1, y_2, \cdots, y_n) \]  

is a sequence, \( \overline{y} = (\overline{y}_1, \overline{y}_2, \cdots, \overline{y}_{n-k+1}) \) a stream called \( y \), in \( \overline{y} = (\overline{y}_1, \overline{y}_2, \cdots, \overline{y}_{n-k+1}) \) is a short sequence of length \( k \) intercepted by \( y \). In addition, since the length of \( \overline{y} \) is determined by \( k \), \( \overline{y} \) is also called a stream of length \( k \) generated from \( y \).
\( r(i) \) represents the length of the \( i \)-th short sequence, and \( e(i) \) represents the frequency weight of the \( i \)-th short sequence, where \( i=1, 2, \ldots, W \), \( r(1) < r(2) \cdots < r(W) \).

Obtain preprocessed data, and generate \( W \) short sequences of different lengths. Generate sequence \( s=(s_1, s_2, \ldots, s_r) \) of \( W \) streams of different lengths: \( \mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^W \), \( \mathbf{s}^i=(\mathbf{s}_{i,1}, \mathbf{s}_{i,2}, \ldots, \mathbf{s}_{i,r(i)}) \) is a stream of length \( l(i) \), collected and preprocessed data will eventually form a sequence, by \( s=(s_1, s_2, \ldots, s_l) \) means, where \( l \) means the number of SQL query fields in the sequence, and \( s_j' \) \( (1 \leq j' \leq l) \) means the \( j' \) field in the sequence \( s \).

Extract different short sequences in each stream and calculate their frequency of occurrence. In the data stream \( s_i \), different short sequence types can be expressed as \( t_1, t_2, \ldots, t_{m(i)} \), \( m(i) \) is the number of different short sequence types in the data stream \( \mathbf{s}^i \). The frequency of each short sequence is calculated according to \( \#i \).

Combine all the different short sequences in all data streams into the normal sequence library, and calculate their weighted frequency of occurrence in the normal sequence library.

Combining all the different short sequences into the normal sequence library POSTGRESQL, the normal sequence library can be expressed as \( Y = (y_1^1, y_2^1, \ldots, y_{m(1)}^1, y_1^2, y_2^2, \ldots, y_{m(2)}^2, \ldots, y_1^W, y_2^W, \ldots, y_{m(W)}^W) \) and the calculation method of the weighted frequency of occurrence of various short sequences in POSTGRESQL is shown in the formula.

The weighted frequency of \( t^i \) in POSTGRESQL is \( \dot{F}_i^i \),

\[
\dot{F}_i^i(i) = e(i)f^i
\] (7)

Sort according to the weighted frequency of each short sequence, and divide them into N-1 groups.

After sorting all the different kinds of short sequences in descending order according to the weighted frequency of occurrence, the result can be expressed as \( Y=(y_1^*, y_2^*, \ldots, y_H^*) \), where \( H=m(1)+m(2)+\cdots+m(W) \). N represents the number of states, \( u \) represents the largest integer not greater than \( H/(N-1) \), and \( v = u + 1 \). The first \( hv \) sequences in \( T \) are divided into \( h \) groups, with \( v \) sequences in each group. The remaining \( (N-h)u \) sequences are divided into \( N-h-1 \) groups, with \( u \) sequences in each group. Therefore, all kinds of short sequences can be divided into \( N-1 \) groups according to the weighted frequency of occurrence, \( \Gamma_1 = \{y_{1,1}^*, \ldots, y_{1,v}^*\}, \ldots, \Gamma_h = \{y_{h,1}^*, \ldots, y_{h,v}^*\}, \Gamma_{h+1} = \{y_{h+1,1}^*, \ldots, y_{h+1,v}^*\}, \ldots, \Gamma_{N-1} = \{y_{N-1,1}^*, \ldots, y_{N-1,v}^*\} \).

2) Behavior feature extraction module based on POSTGRESQL. Before operating abnormal behavior, the warehouse management system needs to establish the behavior characteristics of the normal behavior of the elevated warehouse. In this paper, a discrete-time Markov chain is constructed using the states of short sequences of different lengths as the behavioral characteristics of the elevated library. In the method in this paper, the number of states \( N \) needs to be set in advance. First, it is necessary to obtain the behavior mode of each command in the learning data \( s=(s_1, s_2, \ldots, s_r) \); then according to the N-1 grouping in POSTGRESQL, the Markov chain state of each behavior mode is
generated, and thus obtained a Markov chain state set; finally the transition probability matrix P of the Markov chain is established.

3.3. Abnormal operation

Once the behavioral characteristics of the high-bay warehouse are formed, the basic operation of the abnormal operation warehouse management system is to compare the operation sequence with historical data, and determine whether the two belong to the same high-bay warehouse. The deviation between the behavioral characteristics of the operated elevated library and the historical data is an important indicator of intrusion. Therefore, it is assumed that the historical data belong to the same elevated library. In the abnormal operation stage of the warehouse management system in this article, the classification value determined by the transition probability matrix is the only measure to determine the abnormal behavior, which will greatly reduce the computational burden of the warehouse management system. The main steps are as follows:

1) Obtain and preprocess audit data. In the abnormal operation stage, it is necessary to collect the original SQL command lines generated by the elevated library to be operated (that is, the data that needs to be audited), and the collected data is preprocessed to form a sequence. $c = (c_1, c_2, ..., c_t)$ represents the preprocessed operation sequence, where $c_j$ is the j-th command in the operation sequence, and $t$ represents the number of commands in the operation sequence.

2) Define the state of the operation sequence and generate the state sequence. According to the behavior pattern extraction algorithm and the Markov chain state generation algorithm, taking the normal sequence library $T$ and the operation sequence $c = (c_1, c_2, ..., c_t)$ as input, the state sequence corresponding to the operation sequence $q = (q_1, q_2, ..., q_M)$. The short sequence intercepted by $c = (c_1, c_2, ..., c_t)$ may not be included in POSTGRESQL, so in the state sequence $q = (q_1, q_2, ..., q_M)$, the state value may be N.

3) Obtain the classification value corresponding to the state sequence. In the state sequence $q = (q_1, q_2, ..., q_M)$, for each state $q_i$ $(1 \leq i \leq M - 1)$, the transition probability from $q_i$ to $q_{i+1}$ is $Pr(q_{i+1}, q_i) = q_i, q_{i+1}$. The classification value transferred from $q_i$ to $q_{i+1}$ is defined as,

$$\psi(i) = \begin{cases} 1 & (Pr(q_{i+1}, q_i) > \delta) \\ 0 & (Pr(q_{i+1}, q_i) \leq \delta) \end{cases} \quad (8)$$

In the formula: $\psi(i)$ is the classification value transferred from $q_i$ to $q_{i+1}$; $\delta$ is the critical value of probability, which needs to be set in advance. When $\psi(i) = 1$, it represents a normal transition; otherwise, it represents an abnormal transition has occurred. After calculating the classification values of all states, a sequence of classification values ($\psi(1), \psi(2), ..., \psi(M-1)$) is obtained.

4) Calculate the judgment value. Although individual transfers may deviate from historical precedents, the large number of behavioral collections in the effective elevated library should conform to historical behaviors on the subject, and the intruder's behavior will deviate significantly from the historical behavior. Therefore, the warehouse management system in this paper applies a sliding window to the classification value sequence to obtain a large number of behaviors, and then calculates its judgment value. In this paper, a mean sliding window is used. The judgment value of $\psi(n)$ in the classification value sequence ($\psi(1), \psi(2), ..., \psi(M-1)$) is defined as

$$D(n) = \frac{1}{w} \sum_{m=n-w+1}^{n} \psi(m) \quad (9)$$

In the formula: $D(n)$ is the judgment value corresponding to the classification value $\psi(n)$, $n$ is successively increased by 1 ($w \leq n \leq M - 1$); $w$ is the window size.

5) Determine whether the elevated warehouse behavior is abnormal. The threshold $\lambda$ for judging whether the behavior of the elevated warehouse is abnormal needs to be set in advance according to the system requirements. Whether the behavior of the currently operated elevated warehouse is abnormal is
determined by $D(n)$ and $\lambda$. If $D(n) \geq \lambda$, the current behavior of the operated high-bay warehouse is considered normal, otherwise, if $D(n) < \lambda$, then the current behavior of the operated high-bay warehouse is considered abnormal. In this article, the current behavior corresponds to w classification values $\psi(n-w+1), \psi(n-w+2), \ldots, \psi(n)$, and $w+1$ states ($q_{n-w+1}, q_{n-w+2}, \ldots, q_{n}$) related. The threshold $\lambda$ is a parameter with high sensitivity. The higher threshold, the higher the success rate of abnormal operations and the higher the false alarm rate; the lower the threshold, the lower the success rate of abnormal operations and the lower the false alarm rate.

4. Demonstration of results
Virtual inventory characteristic (VIC), In the signal operation theory, it is used to express the trade-off between the operation rate and the false alarm rate. Figure 1 shows the judgment value curves of the three high-bay warehouses. It can be seen from Figure 1 that the curves of the high-bay warehouse 2 and the high-bay warehouse 3 can be clearly distinguished from the curve of the high-bay warehouse 1, that is, different elevated warehouses. The behavior characteristics of the library are different. Once there is an internal attack on the database, there must be a big difference between the behavioral pattern of the attacker and the behavioral pattern of the normal inventory. The judgment value curve of the operated high-bay warehouse will fluctuate greatly. When the warehouse management system operates to fluctuate when the set threshold is exceeded, it can be considered that there is an abnormality in the inventory.

5. Conclusion
The demonstration shows that the operational performance of the warehouse management system has a great relationship with the selected threshold $\lambda$. The larger the $\lambda$ is, the higher the operation rate and false alarm rate of the warehouse management system will be, and the lower the false alarm rate. Therefore, it is necessary to select a more appropriate value of $\lambda$ through testing. In addition, the number of states...
N will also have a certain impact on the performance of the warehouse management system. Figure 3 shows the curve of the warehouse management system when N is equal to 4, 5, and 6, and the threshold \( \lambda \) is different. It can be seen from Figure 3 that when \( N=5 \), the performance of the warehouse management system is better than when \( N=6 \) and \( N=4 \), and when \( \lambda=0.84 \), the performance of the warehouse management system is optimal.

References
[1] Lü L, Chen DB, Zhou T. The small world yields the most effective information spreading [J]. New Journal of Physics, 2011, 13: 123005.
[2] W Zikopoulos P, Eaton C. Understanding Big Data: Analytics for Enterprise Class Hadoop andStreaming Data. New York: McGraw-Hill Osborne Media, 2011.1-10.
[3] SUN Y, WANG X, TANG X. Deep learning face representation from predicting 10, 000 classes [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway,NJ: IEEE, 2014: 1891-1898.
[4] Mccormick B H, DeFanti T A, Brown M D. Visualization in Scientific Computing[J]. IEEE Computer Graphics and Applications, 1987, 7(10): 69.
[5] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107-113.
[6] N. Wang, X. Li, X. H. Chen. Fast three-dimensional Otsu thresholding with shuffled frog-leaping algorithm[J]. Pattern Recognition Letters, 2010, 31(13): 1809-1815.
[7] S. S. Herag u, X. Cai, A. Krishnamurthy, et al. Analytical models for analysis of automated warehouse material handling systems [J]. International Journal of Production Research, 2011, 49(22): 6833-6861.