ATLANTIS PRESS
Advances in Computer Science Research (ACSR), volume 76
7th International Conference on Education, Management, Information and Mechanical Engineering (EMIM 2017)

Research on the Construction Method of the Frequency Pattern of Populations in Cities

Huanliang Sun[1, a], Zhiqing Zhang[1, b] and Junling Liu[1, c *]

[1]School of Information & Control engineering, Shenyang Jianzhu University, Shenyang, China

[a]1968207285@qq.com, [b]1293804902@qq.com, [c]2428872271@qq.com

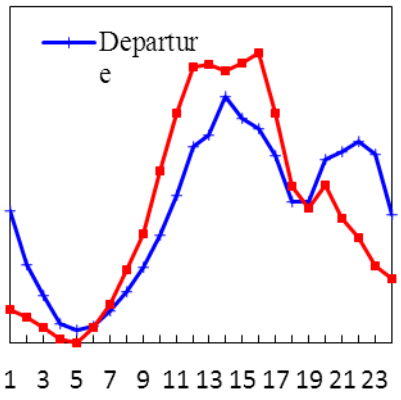
*The Corresponding author

Keywords: Frequent pattern; Hierarchical clustering tree; Branch skew; Abnormal sequence

Abstract. In recent years, with the popularity of mobile terminal equipment and the arrival of the Internet era, people in the city use mobile phones more and more frequently. And the wide use of mobile phones has produced a large number of saptio-temporal data indicating the location as well as containing the law of population mobility in cities, which can be applied to transportation, commercial location, social networking applications, crime analysis, and meteorological research and so on. And if you would like discover this law, you need to carry out complex processing of these data transformation. This paper takes this as the starting point, and mainly studies the construction method of the frequent pattern of Population in cities. In the process of studying the construction method, this paper proposes a method to discretize and serialize the data according to the theory of urban Movement of Population and the existing data mining theory, and then use the hierarchical clustering number for these data Algorithm, as well as use the concept of branch skew to remove the abnormal sequence, consequence generating the pattern set. At the end of this paper, we use the real data set to test the construction method. The results show that the model set generated by this method is a good reflection of the law of population mobility in the city.

Introduction

With the popularity of mobile devices such as mobile phones and the advent of the Internet era, mass data come into being, which include different sizes and types[1]. The great quantity, diversity, authenticity and low value density of data indicate itself being typical big data. Applying to data analysis technology, we can obtain a great deal of interesting laws digging from these data[2]. The statistical data in Fig. 1.1 shows the record of getting on and off the taxis in two regions of Shenyang .



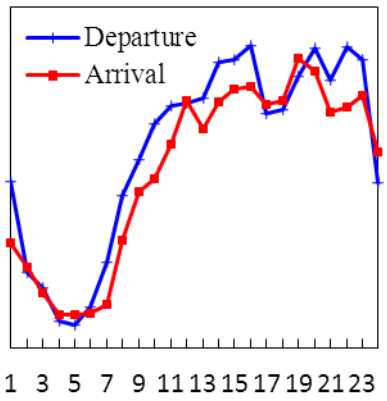


Figure 1. Example of travel law (left) Figure 1.1 Example of travel law (right)

Fig. 1.1 shows more arrival in the evening and more departure in the morning, on the contrary more arrival in the morning and more departure in the evening [3-5]. From the above rules we can infer that the left graph is similarly reflecting departure and arrival of uptown, while the right graph

is similarly reflecting departure and arrival of business district or campus [6-7]. In this paper, we call this law the frequent pattern of Population, which can be used to predict the crowd flow, commercial location and police dispatch. Based on the above demand, this paper proposes a model of urban population movement, actually obtaining the pattern of population movement from making use of the location data of urban population mobility, apply to predict the flow of urban population [8-9].

Construction Model of Flow Pattern

Design Ideas of Model. Step ①: This step converts the data of the acquired group object activity into a four-tuple representation, as a consequence of discretizing the data.

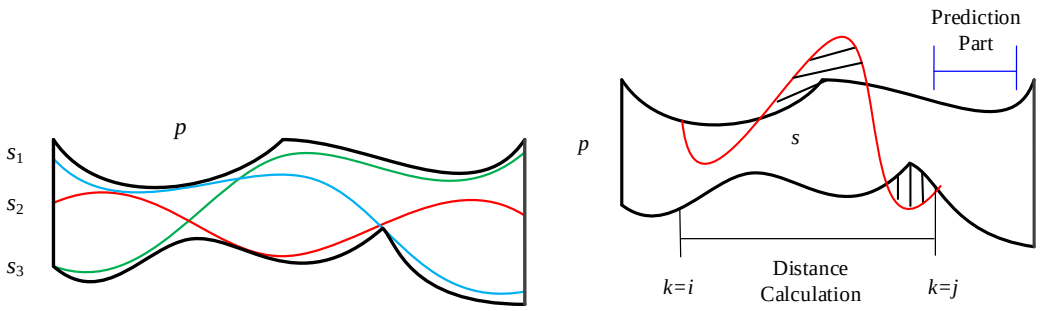
Step ②: Select the area, and the data in the region are bin processed in chronological order, respectively make a subtraction on the departure and arrival data for every group object in a bin to get the flow sequence and complete the serialization of data, and then use the natural day division method to further subdivide these sequences with the aim of breaking up the obtained data into day sessions, so that the next hierarchical clustering algorithm can be normally performed.

Step ③: The data is divided into two parts: the training data set and the test data set. The training data set is used to generate the model. The test data set is used to verify the prediction accuracy of the model. Step 3 mainly carries on the operation to the training data set, according to the hierarchical clustering algorithm on the serialized data to generate the skew level clustering tree.

Step ④-⑤: Introduce the concept of skew and two theorems to apply the clustering automatic selection algorithm to remove the local anomaly sequence and generate the pattern set.

Step ⑥: Evaluate the generated pattern and finally get the model set after the test.

Related Definitions. Definition 1: Pattern. The pattern is a summary description of the subset $S_p = \{S_{p1}, S_{p2}, S_{p31}, \cdots, S_{pm}\}$ of the sequence set S in a particular spatial region D. The pattern is shown by the wedge with the upper and lower bounds and with length of m, expressed as $\{[U_1, L_1], [U_2, L_2], [U_3, L_3], \cdots [U_m, L_m]\}$. Pattern error is $\delta = |U_1 - L_1| + |U_2 - L_2| + |U_3 - L_3| + \cdots + |U_m - L_m|$.



(a) Creating pattern p by s1, s2, s3 (b) Computing d(s, p)

Figure 2. Pattern and pattern spacing

Definition 2: The distance between the sequence and the pattern. If the fact which the length of mode S are equal to that of sequence P and the time span of the both are in coincidence exists, the partial overlap of the dimension is set as a interval from the i-dimensional to j-dimensional , thus the distance between the sequence S and the mode P can be expressed as $d(s, p) = \sum_{k=i}^{j} d_k$ .There are three cases about the calculation of each dimension of d_k : (1) when $p \cdot L_k \le s \cdot x_k \le p \cdot U_k$, obtain $d_k = 0$; (2) when $s \cdot x_k > p \cdot U_k$, obtain $d_k = |s \cdot x_k - p \cdot U_k|$; (3) when $s \cdot x_k < pL_k$, obtain $d_k = |s \cdot x_k - pL_k|$. The overlapped shaded portion shown as Figure 2 (b) is the

sum of the nonzero distance of the sequence S and the pattern P, and the rest is the distance whose value is zero between the sequence S and the pattern P.

Definition 3: Pattern set. Pattern set is a collection of patterns in definition 1. The pattern set $P=\{p_1, p_2, p_3, \cdots p_r\}$, which contains r patterns, is a summary description of every of the flow sequence set S.

Definition 4: Predictive error. There is a certain error in the predicting sequences by the mode of the pattern set, and the error is defined as

$$Error=\frac{\sum_{k=j+1}^{m} d_k+\sum_{k=j+1}^{m}\left(U_k+L_k\right)}{|m-j|}$$

In which $\sum_{k=j+1}^{m} d_k$ is the distance between the predicted sequence S_K' of the sequence S and the pattern, $\sum_{k=j+1}^{m}\left(U_k+L_k\right)$ is the mode error of the pattern and the length of prediction sequence S_K' is $|m-j|$.

Definition 5: Branch skews. Define a node p_i, then the branch skew of p_i is the sum of the height suntraction between the node and the left and right subtree of all descendants of it , which can be expressed as $\sigma\left(p_i\right)=\sum_{j=0}^{1}\left(H\left(p_j, Lchild\right)-H\left(p_j, Rchild\right)\right)$, in which p_j is the descendant note of p_i, and $H\left(p_j, Lchild\right)$ and $H\left(p_j, Rchild\right)$ are about the height of the subtree of p_i.

Maximum skew theorem of the tree. A skew-level clustering tree with n leaf nodes has a maximum deflection of $\frac{(n-1)(n-2)}{2}$.

Complete binary tree theorem. A skew clustering tree with n leaf nodes, if the tree is a complete binary tree, and the number of leaf nodes n satisfies $2^m<n\leq 2^{m+1}$, then the range of the tree skew is $[0,\ m]$.

Skew Hierarchy Clustering Tree. Hierarchical clustering tree at the time of establishment will select the smallest sequence or pattern each time to merger, that is, the increment of the sum of squares of each dimension of a sequence or pattern, it is effective in removing the abnormal sequence while achieving automatic selection mode. In this paper, the combination of the sequence or mode is used in way of high left low right, that is, the high subtree as the left one, the lower subtree as the right one. A four-tuple $\left(n_p, \delta, \sigma, h\right)$ is stored in the node p, which includes the number of sequences n_p, the mode error δ, the branch skewness σ, and the branch height subtraction h.

Input of Algorithm 1 is the sequence set, and its output is the hierarchical clustering tree. Step (1) put the sequence in the input sequence set S into the candidate set M; Step (2) counts the number of sequences contained in the candidate set M and judges whether the number of sequences is greater than 1. If the number is greater than 1, step (3) - step (12) is performed; Step (4) calculates the distance between every two of all the sequences in the candidate set M. The distance between the sequences is calculated by the substraction between the upper and lower bounds of the dimension. The distance between the sequence and the pattern or between pattern and pattern is calculated by the formula described in section 3. Step (5) - Step (7) Calculate the distance of the two sequences according to step (4), find the two sequences Mi and Mj with the smallest distance, and merge the two sequences into a new mode Mij, at the same time, the model number np, the model error δ, the branch skewness σ, and the branch height substraction h included in the newly generated pattern

Mij are calculated to generate; Step (8) - Step (9) The two sequences Mi and Mj found in step (5) are removed from the candidate set M; step (10) adds the pattern Mij merged in step (6) to the candidate set M; Step (11) The pattern Mij merged and generated by the sequence Mi and Mj in step (6) is added to the hierarchical clustering tree T; step (13) It is not cycled until when there is only one sequence in the candidate set M, the hierarchical clustering tree T is also established.

Cluster Automatic Selection Algorithm. After the hierarchical clustering tree being generated and splitting it, the cluster can be obtained. On the basis of the skew, a method is used in this paper which can automatically select the sequences number in the balance cluster to remove the anomalies, as shown in Algorithm 2.

The input of algorithm 2 is the hierarchical clustering tree T and the least number of sequences $n_{\min}$ in each pattern. Step (1) Calculate the actual skewness $\sigma(T)$ of T and the maximum skewness generated by np nodes (Theorem 1); Step (2) Calculate the number of abnormal sequences $n_0 = \omega \cdot \frac{\sigma(T)}{\sigma_{\max}} \cdot n_p$ to be removed in T; Step (3) store the nodes in the list L in descending order of σ; Step (4) initialize variable i to be 0, which is used to store abnomal sequences number, ; Step (5) - (10) Traverse the nodes in the list L, remove the exception sequence, each time start from the node with the largest σ value, delete the node's right subtree until the number of leaves to be eliminated satisfies the condition $i \geq n_0$, then stop the cycle; Step (11) estimates the maximum value $n_{\max}$ of the pattern length according to the actual skewness $\sigma(T)$ of T; Step (12) - Step (16) Uses the maximum Heap to complete the access to the node, and the node containing the number of sequences is preferentially split and handle. When the number of sequences contained in the node is larger than the upper limit $n_{\max}$ of the mode, The left and right subtrees are placed in the heap, and the abnormal sequence is processed during the process of split, and the set of patterns is output when the number of nodes satisfies $n_{\min} \leq |p| \leq n_{\max}$.

Experimental Evaluations

Experimental Environment and Data Sources. The programming language of the urban Movement of Population model proposed in this paper is JAVA, and the experimental evaluation platform is four 64 GHz machines with 3.2 GHz Core (TM) i5 CPU and 4.0GB memory. Experimental data is the record of getting on and off of 11,000 taxis in Shenyang , about 100,000 records each day, and the total data size is 1.9G.

Pattern of Visualization Area. The experiment selected five characteristic areas of Shenyang city as the research object, respectively, Olympic Sports Center, Yucai Campus, Zhongjie Commercial Area, Shenyang North Station and Xingshun Night Market. These five areas belong to large-scale conference hall, campus, large commercial area, large passenger terminal and large Pedestrian Street. During the experiment, the area of each area is calculated according to Table 1.

Table 1 Area of each area

Region name	**Default size/** km^2
Olympic Sports Center	1.41
Yucai Campus	3.79
Zhongjie Commercial Area	2.15
Shenyang North Station	3.79
Xingshun Night Market	2.15

The population pattern of the experimental area is constructed and the visualization pattern set is generated, as shown in Fig. 3.1, which shows the Frequent pattern set of Xingshun Night Market and Yucai Campus. It can be seen from the definition of Chapter 3 that a pattern set is composed of multiple patterns, in which each pattern is a wedge surroundded by the upper and lower bounds. In order to increase the expression effect of the results graph, the upper and lower bounds are described by different colors.

From the figure, it can be clearly seen, in Yucai campus, the arrival and departure in morning, noon and evening are more than that in other times, and in Xingshu night market, arrival in the evening is more than departure, while departure is more than arrival in early morning, which more intuitively reflects the law of the flow pattern in the areas, and the law of the flow pattern is consistent with the law of life corresponding to the characteristic area, which further validates the validity of the algorithm proposed in this paper.

Conclusions

This paper studies the construction method of the frequent pattern of Population in cities, designs the construction model of Frequent pattern, discretizes the data, counts the number of out-in times periodically to serialize the data. The serialized data is clustered using the hierarchical clustering method, and the proposed skewness can be used to remove effectively the local anomaly sequence and balance the difference of pattern. Using the authentic data set to verify the construction method, it is proved that the proposed method has high prediction accuracy.

Acknowledgement

This work was supported by the Liaoning Provincial Social Planning Fund (L15BGL017) and science and technology program of Liaoning(20170540767).

References

[1] Shao J I, Zhu Z S. Study of Population Movements and the Changes in the Pattern of Urbanization[J]. Research on Economics & Management, 2013.
[2] Cao C. The Study of County Urbanization Path Mode Based on the Influence of Seasonal Population Movements: A Case of Jinzhai County[J]. Urban Development Studies, 2013.
[3] Benjamin. Segregation and Urban Form: Towards an Understanding of Dynamics Between Race, Population Movement, and the Built Environment of American Cities[J]. 2014.
[4] Korpi M. Migration, wage inequality, and the urban hierarchy : empirical studies in international and domestic population movements, wage dispersion and income: Sweden, 1993-2003[J]. Economic Research Institute Stockholm School of Economics, 2011, 25(1):15-25.
[5] Jun M A. Influence of Rural Population Movement on Rural Development under the Background of Urbanization[J]. Journal of Anhui Agricultural Sciences, 2012.
[6] University R. Introduction to Population, Urbanization, and the Environment[J]. 2014.
[7] Zhang J, Wu L. Influence of human population movements on urban climate of Beijing during the Chinese New Year holiday[J]. Sci Rep, 2017.
[8] Walaszek M. Availability and Organization of Social Services in the Areas of Intense Urbanization (Example of Poznan Agglomeration)[J]. Acta Universitatis Nicolai Copernici Ekonomia, 2013, 44:93-112.
[9] Wu L, Zhang J. Assessing population movement impacts on urban heat island of Beijing during the Chinese New Year holiday: effects of meteorological conditions[J]. Theoretical & Applied Climatology, 2017:1-8.