http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 11 (3): 334-338, 2012
ISSN 1812-5638 / DOI: 10.3923/itj.2012.334.338

The Research on Traffic Flow Pattern Clustering Based on Frequent Sequences Similarity

[1]Jin Du, [1]Xiangmo Zhao, [2]Jun Hao and [1]Haiwei Fan
[1]School of Information Engineering, Chang'an University, Xi'an 710064, China
[2]Bureau of Xi'an Rail-way, China

Abstract: The research on traffic flow has shown that certain latent laws of traffic status evolvement must exist among traffic data. However, little work has been done to support this view by experimental methods. In order to lay a solid foundation for intelligent transport management, a new approach is presented in this paper to investigate the traffic flow pattern on the basis of traffic sequence mining. Firstly, comparing the similarity between traffic status sequences, here a similarity measure is defined in this approach. Secondly, through sequence similarity calculated, a traffic status sequence clustering algorithm, FSCS (Frequent Sequences Clustering based on Similarity), is further proposed. Lastly, the clustering results were estimated and the optimal clustering pattern is preserved. The experiment with real traffic data revealed that there are four universal traffic flow patterns exist.

Key words: Traffic flow, frequent sequence, sequence clustering, sequence similarity

INTRODUCTION

In the development and application of ITS (Intelligent Transport System), the traffic flow pattern distinguishing is very important because it describe the general characteristic for traffic status transforming. Today, there were various kinds of monitor sensors and cameras were disposed on the roads, these devices collected plenty of traffic condition data in various environments and periods. The traffic status parameters, such as traffic flux, road density, vehicles speed, delay duration, can be analyzed according to those traffic data gathered. These data were existed as a formal of data flow and can not be saved by traditional database management system. Because those raw traffic condition data are high dimensional, high-order, random, change with time, nonlinear, therefore, the work to distinguish the traffic flow status is very hard. Further, to forecast the trend of traffic flow evolvement in short time is very difficult using those data directly.

Recently, the study for traffic flow status distinguishing focus on machine learning or statistical methods. However, as we known, the transformation of traffic flow is random and dynamic, there are strong dependence relationship among different traffic status in different period. It might have some potential rules according to the time sequence. So, the technology of data mining become very popular because it is proved very useful to discover the dynamic rules of traffic flow and will provide more profound theory support for intelligent transport control. Chen (2005) proposed clustering traffic flow into 2, 3 and 6 clusters by fuzzy clustering method, based on the theory of 3-items flow, Wang and Weng (2009) adopt fuzzy C-means clustering method and divide fast road traffic flow status in several classes. Wang and Chen (2009) explored to divided traffic status into 2 classes through incremental Bayes classing algorithm. However, all the work mentioned were emphasis on the static example data and rely on original training data too much to describe the real traffic flow change regulation.

The concept of sequence pattern was proposed by Agrawal (Guo *et al.,* 2004) and Strikant (Xing and Shen, 2004): giving a set of different sequences which consisted of different elements by some order. The elements were composed with some items. Through study this kind of sequence data set, we can draw some unknown rules which we call as sequence patterns.

In this paper, a method based on frequency sequence clustering was proposed. Firstly, the status of traffic flow were observed and marked with several status labels, Secondly, the traffic status data were dealt into traffic flow sequences. Through sequence similarity calculating, the traffic flow was clustered into several traffic flow patterns according to the distance between different traffic flow sequences. These patterns could describe the general rules hided in real time traffic data.

The sequence clustering had been investigated in recent years. The Microsoft Sequence Clustering algorithm (Duan *et al.,* 2006) is a sequence analysis algorithm provided by Microsoft SQL Server 2005 Analysis Services (SSAS), this algorithm finds the most

Corresponding Author: Jin Du, School of Information Engineering, Chang'an University, Xi'an 710064, China

common sequences by grouping, or clustering, identical sequences together. Jiangjiao Duan proposed a hidden markov model-based hierarchical time-series clustering algorithm (HBHCTS), the algorithm modeling time sequence based on hidden markov model, get the initial model sets sequences on the basis of "most similar", moreover, clustering these model sets by updating and combining iteratively (Halkidi and Varzirgiannis, 2001).

The calculation of similarity between traffic flow sequences is very important for traffic flow pattern discovering based on sequence clustering. It is reported that the similarity between status sequence sets can be calculated as (Xu *et al.*, 2006):

$$Sim(as_i, as_j) = \frac{|as_i \cap as_j|}{|as_i \cup as_j|} \quad (1)$$

However, the similarity calculation just regarded the ratio that same status occurred in both two status sets and do not concern the occurred order. In fact, the status sets (a,b) and (b,a) are not same sequence obviously.

TRAFFIC FLOW SEQUENCE PATTERN

Definition 1: Traffic status: The Traffic Status records the current situation of a site at some moment for someone road, it is represented as following tuple:

$$a = (RID, TID, Situation, Time, Delay) \quad (2)$$

where, RID is the unique identity of a road site, the TID identify a unique status, Situation represents the semantic information by the name of current traffic status, the Time is the time stamp of current traffic status and indicates the time when occurred, the Delay indicates the duration of this status.

Definition 2: Traffic status sequence: The traffic status sequence is a set of statuses which occurs in one section of traffic events and are order by their time stamp, represented as:

$$a = (a_1, a_2, ...a_i) \quad (3)$$

Definition 3: Traffic flow sequence pattern: The Traffic Flow Sequence Pattern is the formalized representation of traffic flow characters and it describes the potential regularities hided in traffic flow sequences.

The traffic flow sequence pattern can be obtained by traditional data mining or statistical methods. Statistical method could get the omnibus and remarkable traffic rules such as sum of time of some status, occur frequency, total sites the status occurred and so on. However, dynamic features of traffic flow can not reflected by statistic method. In another word, the potential rules of traffic flow may be omitted. In this paper, the traffic flow sequence pattern is analyzed by frequent sequence mining and the traffic flow pattern can be represented in form of frequent traffic status sequences pattern:

$$TP = \{(as_1, spt_1), (as_2, spt_2)...(as_n, spt_n)\} \quad (4)$$

where, the as_i is a frequent traffic status sequence and the spt_i is the frequency of as_i occurred .

TRAFFIC STATUS SEQUENCE CLUSTERING

Definition 4: Traffic status sequences similarity: Supposing there are two traffic status sequences, as_i and as_j, where:

$$as_i = (a_{i1}, a_{i2},...a_{in}) \text{ and} (t.,t_{j2},...t_{jm})$$

The similarity between two traffic status sequences can be calculated in three steps.

Step 1: Longer sequence compressing: Without loss of generality, supposing $n \leq m$(as_i is not longer than as_j) and all of statuses which belong to as_j and do not appear in as_i should be deleted from as_j. For example, as_i = (a,b,c) and as_j = (a,d,b,e,b,c). After compression, as_j is transformed in to as_i ' = (a,b,b,c) with the compressing rate $|as_i'|/|as_j|$. The size of as_j ' is marked as $|as_j'|$ = t. The aim of compression is to keep the order of all elements of as_i occurred in as_j and the atomic of as_i is not regard.

Step 2: Sliding window comparison: The similarity between as_i and as_j was compared by means of slide windows method. Here, the size of slide window equal to the length of the shorter sequence between as_i and as_j and marked as $size_w$ = min (n,t). Again without loss of generality, supposing $n \leq t$ (as_i is not longer than as_j'), then as_i is taken as target sequence and $size_w$ = n. Shown as Fig. 1 b, at Tth, as_i is compared with sub-sequence of as_j which right fill in slider window (marked as sT).

As the Fig. 1 a shown, the initial comparison start from the tail of as_i and head of as_j '. In Fig. 1 c, the final comparison ended at the head of as_i and tail of as_j '. Therefore, a special sequence as_h = (0,0,...0) is added as a prefix and postfix to as_j'.

Where, $|as_h|$ = n-1 and any element of as_h do not belong to as_i ($a_{sh} \cap a_{si} = \varnothing$). Set the slide step to 1, totally t+n-1 comparisons are completed between as_i and s_T as the window sliding.

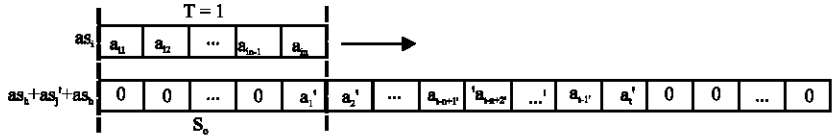

Fig. 1a: The comparison at being (T = 1)

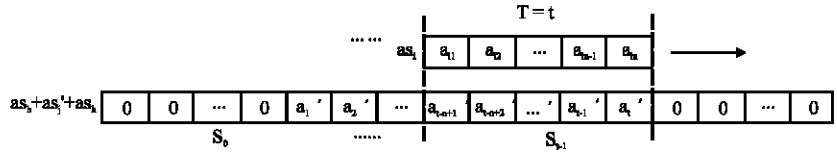

Fig. 1b: The comparison at t moment (T = t)

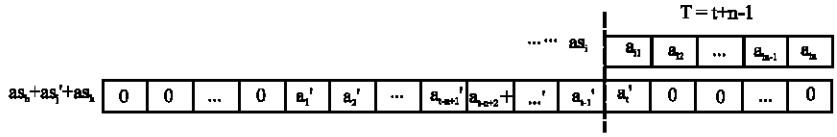

Fig. 1c: The comparison at end (T = t + n-1)

At each comparison, the similarity between as_i and s_T is calculated as:

$$S(as_i, s_T) = \sum_{k=1}^{n} s(as_i(k), s_T(k)) \quad (5)$$

where, as_i (k) is the kth element in as_i and s_T (k) is the kth element in s_T. If as_i (k) s_T (k), then s (as_i (k), s_T(k)), else s (as_i (k), s_T(k)).

Step 3: Similarity calculating: According to above illustration of slide window comparisons between as_i and as_j, the similarity calculation about as_i and as_j is processed as following.

First, to calculated the sum of similarity between as_i and s_T and make out the similarity between as_i and as_j' evenly.

$$\bar{S}(as_i, as_j') = \frac{\sum_{T=1}^{t+n-1} S(as_i, s_T)}{t+n-1} \quad (6)$$

Secondly, the mean similarity $\bar{S}(as_i, as_j')$ as_i is multiplied with the ratio of compressing from as_i to as_j', the final result is gained.

There for, the similarity between as_i and as_j can be calculated by formula:

$$Sim(as_i, as_j) = \frac{\sum_{T=1}^{t+n-1} S(as_i, s_T)}{t+n-1} \cdot \frac{t}{m} \quad (7)$$

As the hypothesis testing that intra similarities are very different from those of other group else, the following profound research will be pushed forward. This paper put forward a novel clustering algorithm named as FSCS (Frequent Sequences Clustering based on Similarity) which based on the similarity analysis between traffic status sequences, by this algorithm, the data set of traffic status sequence will be partitioned into several traffic status sequence clusters which are taken as traffic flow patterns.

The process of FSCS is represented as following:

Input: D = (as_i, as_2,... as_n), k
Output: k clusters (c1,c2,…ck)
Algorithm:
Initializing_Clusters();
D ' = D-{cI_o, $c2_o$,...cK_o}
For i=1 to |D'| do
Begin
For j =1 to k do
$V_{j,j}$ = Sb({cj_{i-1}, as '$_1$)}-Sb (cj_{i-1})
If $v_{j,j}$ is Max (v_{1j}, $v_{2,i}$,... $v_{k,i}$)
Thencj '$_1$ = cj_{i-1} + {as '$_1$}
//add as '$_i$ to cj '$_{i-1}$ whose Sb raised most significantly with as '$_i$ added to
End.

In procedure of Initializing_Clusters(), 2k traffic status sequences are selected from D and assembled into k clusters randomly. The initial clustering pattern marked as C_o = ($c1_o$, $c2_o$,...ck_o) and each ci_o is composed of 2 sequences.

Because the distribution of traffic status sequences is unpredictable, an effective clustering pattern assessment method should be adopted to estimate the quality of clustering result and detect whether selection

of parameter is optimal. Therefore, the traffic status clustering must be an iterative process.

Here, the quality of clustering result is estimated according to the intra-cluster similarity and inter-cluster similarity. With the input parameter k is adjusted and the initial clusters (C0) varied, a new clustering result (Cp_k = (c1, c2,...ck)) is generated. To calculate the ratio (marked as σ) of the mean intra-cluster similarity (marked as Sb (CP *k*)) to the mean inter-cluster similarity (marked as Sw (CP *k*)), the max(σ *k*) indicates that the optimal clustering pattern is obtained.

Where:

$$\sigma_k = \frac{\overline{Sb}(CP_k)}{\overline{Sw}(CP_k)} = \frac{\frac{1}{k}\cdot\sum_{i=1}^{k} Sb(ck_i)}{\frac{1}{C_k^2}\cdot\left(\sum_{i=1}^{i=k}\sum_{j=i+1}^{j=k} Sw(ck_i, ck_j)\right)} \quad (8)$$

EXPERIMENT AND APPLICATION

The original experiment data were collected through Chang'an University's road monitoring cameras set on Xi'an south 2nd loop road for 3 months. After data preprocessing and labeling, we get 379 traffic status sequences. All the sequences were assembled into one data set *D* and the algorithm FSCS was adopt to divide *D* into several traffic flow clusters. Here *k* was set from 4 to 20 and the seed start from 0 to 140, totally 17*141 clustering results were generated. For each *k*, by comparing the average intra-cluster similarity of traffic flow patterns, the optical seed is found out.

Figure 2 shows the curve of average intra-cluster similarity varied with k. Obviously, when k = 4 or 5, the traffic flow clustering pattern is optimal.

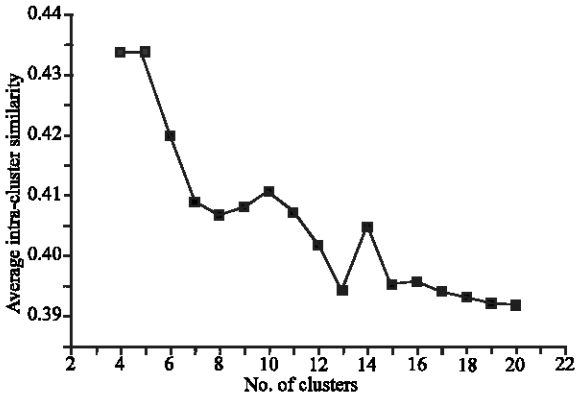

Fig. 2: Average intra-cluster similarity with respect to number of clusters

CONCLUSION

The research of traffic flow similarity among traffic status sequence can provide theory support for building an intelligent transport system. In this paper, a novel method for traffic status sequence clustering based on sequence similarity is proposed. According to calculations of traffic flow similarity, the characters of traffic flow evolvement hide in random traffic data were investigated. By using the sequence clustering algorithm FCSC proposed in this paper, those traffic status sequences were clustered into four traffic flow patterns.

The work presented in this paper is just the first exploration for traffic flow by sequence clustering technologies. It is necessary that some methods and algorithms such as FSCS should be further improved with some optimizing process. In the future, the research result should be applied in forecast for traffic flow in short time by the traffic flow patterns matching in traffic control system. And the mechanism of traffic guiding based on those patterns should be explored.

ACKNOWLEDGMENT

This research has been supported by the Nature Science Foundation of Shaanxi Province (2009JQ8002), The special Fund for Basic Scientific Research of Central Colleges and the Special Fund of Basic Research Support Program of Chang'an University (CHD2011JC021), shaanxi engineering and technical research center for road and traffic intelligent detection.

REFERENCES

Chen, D.W., 2005. Classification of traffic flow situation of urban freeways based on fuzzy clustering J. Transp. Syst. Eng. Inform. Technol., 1: 62-67.

Duan, J., Y. Xue, Z. Lin, W. Wang and B. Shi, 2006. A novel hidden markov model-based hierarchical time-series clustering algorithm. J. Comput. Res. Dev., 43: 61-67.

Guo, L., X. Xiang and Y.C. Shi, 2004. Use Web usage mining to assist online e-learning assessment. Proceedings of the IEEE International Conference on Advanced Learning Technologies, August 30-September 1, 2004, Beijing, China, pp: 912-913.

Halkidi, M. and M. Varzirgiannis, 2001. Clustering validity assessment: Finding the optimal partitioning of a data set. Proceedings of the IEEE International Conference on Data Mining, November 29-December 2, 2001, San Jose, CA., USA., pp: 187-194.

Wang, D. and X. Chen, 2009. Application of incremental bayes classifier on traffic congestion identification. Comput. Aided Eng., 16: 56-59.

Wang, X. and X. Weng, 2009. Classification of three-phase traffic flow of urban expressway based on fuzzy C-means clustering. J. Transp. Inform. Saf., 1: 149-152.

Xing, D. and J. Shen, 2004. Efficient data mining for web navigation patterns. Inform. Software Technol., 46: 55-63.

Xu, G., Y. Zhang and X. Zhou, 2006. Discovering task-oriented usage pattern for web recommendation. Proc. Aust. Database Conf., 49: 167-174.