Dynamic Estimation of Hospital Reservation Registration Service Time in the Basis of Dual Attribute Similarity

Bo Zhang¹, Dehua Kong², Yongxia Zhao³
School of Computer Science, Wuhan Donghu University, Wuhan 430212, China
*Corresponding author’s e-mail: beata_kong@sina.com

Abstract. Considering the particularity, complexity and uncertainty of influencing factors, a dynamic estimation model of reservation registration service time based on binary attribute similarity in the hospital is proposed to solve the problem of hospital reservation registration service time estimation. The model combines hospital process expert knowledge with machine learning of historical data, and uses binary attribute similarity to construct the model of hospital data object. The results show that the proposed method can effectively estimate the booking service time and have high accuracy.

1. Introduction
With the rapid popularization of smart phones, a variety of hospital applications appear. In the process of hospital development, the difficulty of seeing a doctor not only troubles the hospital, but also makes the patients feel inconvenient. Prediction of hospital appointment registration algorithm based on binary attribute similarity solves the difficulty of registration and payment when patients see a doctor.

At present, the difficulty of registration is widely felt and urgently needed to be solved. Many patients would queue up all night in order to get experts’ consultation. Many ticket scalpers also take this as a way of making profits, which reduces the trust of medical enterprises. At the moment, there are three main ways of registration: telephone reservation registration, on-site registration and Internet registration. However these three ways of registration have defects. Telephone reservation registration has the shortcomings of busy telephone and unclear registration information for doctors, which makes the patients unable to obtain satisfactory registration experience. On-site registration is a nightmare for doctors. On-site queuing takes a lot of time, and the quality and quantity of the number are usually not satisfied. However, the way of Internet registration and reservation needs the equipment that can access the Internet, thus it is not a convenient way of registration. Patients can know the real-time information of each doctor's remaining number and visiting situation according to the situation of experts so that they can determine their own visiting time and they can make appointments through the login account and print the number to the hospital self-help, which greatly improves the efficiency of medical treatment. For the patients who come back for consultation, they can make a direct reservation through self-service, which reduces the time of queuing and enhances the efficiency of registration.

Due to human negligence, equipment anomalies or sampling methods, the original data of reservation registration often have data errors, data missing, inconsistent, duplicate or contradictory situations. If these original data are analyzed directly, it is difficult to get high quality of data analysis results. There are many meaningless components in massive real data, which seriously affects the efficiency of data analysis algorithm.

The original data of reservation registration is collected directly from various platforms such as devices or systems which always comes from heterogeneous data sources. There are a lot of noise,
discrete points and redundancy. Generally speaking, these are incomplete and inconsistent dirty data and the quality of data is too poor to directly carry out reservation registration data mining, or the results of mining are unsatisfactory. In order to determine whether an HTTP request is normal or malicious among the huge amount of data during the check-up of personnel appointment registration, they should be grabbed firstly. A probe program is deployed on a specified gateway thus all HTTP requests passing through the gateway are captured and stored. However the program captures some unknown irregular characters and the processing method firstly decodes them and defines a certain format to store them. At this time, the original data of the reservation registration obtained may not be clean, there will be noise, and there might be some deficiencies.

The missing values are mainly missing or incorrect data, which may include errors in data input by human or computer and misunderstanding errors in input. It is also possible that there are problems with data collection equipment, file conversion and data loss, which must be cleaned before data analysis to reduce the impact on the results of subsequent data analysis.

Delete records of missing attributes directly. A complete data set can be obtained by deleting records with missing information but this method has great limitations. In order to reduce the amount of data in exchange for complete information, it is possible to discard a large number of useful information for data analysis. Manual fill-in method is effective for special data set, but it is not feasible when the amount of data is large. It could use global constants to fill in missing values. Each missing value is replaced by a global constant, but this easily interferes with the analysis results. The missing values are filled with the central trend measure of the attribute.

The original data mining system of reservation registration must include the preprocessing module of reservation registration data, which is task-oriented and guided by relevant professional knowledge. It uses a new "task model" to organize the original data, discards some attributes unrelated to the mining task, fills the missing values, filters the noise points that will affect the results, and provides a complete and consistent data mining algorithm. Accurate and effective data can reduce data processing involved in mining calculation, improve mining efficiency and discover knowledge accurately and qualitatively.

2. Binary Attribute Similarity of Reservation Registration Original Data

Range, also known as full distance, is the difference between the maximum and minimum values of the total observed values of a set of data sets. According to the maximum and minimum values in the data, ignoring other values and the difference between the observed values, range reflects the maximum degree of discreteness in a set of data sets.

All the observed values in the data set are arranged in incremental order, and then the data are divided into continuous sets of basically the same size. Every certain distance to a data point on the data distribution is the quantile. When the data set is divided into K parts, a total of k-1 data points which can be called as K-quantile, can be extracted from the start point to the end point of the data set. Ordered data sets are divided into N parts with equal spacing and each part contains 1/N of data in the data set. At this time, N-1 data points are generated to form N-quantile. The N-quantile gives some hints about the center, distribution and shape of the data distribution. The first N-quantile is Q1, which is in the position of 1/N of data distribution; the second N-quantile is Q2, which is in the position of 2/N of data distribution; the (N-1)th N-quantile is Q_{N-1}, which is in the position of N-1/N of data distribution; the distance between the first N-quantile and the (N-1)th N-quantile is the N-quantile range IQR.

\[ IQR = Q_{N-1} - Q_1 \]

This project uses five-digit data: the first quintile data is distributed in 20%; the second quintile data is distributed in 40%; the third quintile data is distributed in 60%; the fourth quintile data is distributed in 80%, and the distance between the fourth and the first quartile is the quintile range IQR.

Analysis of Data object’s similarity finds that the similarity between objects is a numerical measure of the degree of similarity. Data similarity is also the degree of proximity. The value of proximity is usually non-negative. The range of values is between 0 and 1 that 0 represents the
complete dissimilarity and 1 represents the complete consistency. Computing the similarity is to convert the degree of difference between the calculated objects, that is, the dissimilarity. The more similar the two objects are, the smaller the difference is and the lower the difference is.

### Table 1. Binary Attribute Value Table

| Object i | 0 | 1 | sum |
|----------|---|---|-----|
| 0        | s | q | r+s |
| 1        | 0 | q | q+r |
| 0        | t | s | s+t |
| sum      | r+t| p |     |

There are only two states for binary attributes: 0 and 1 or True and False. The similarity measure of binary attributes can be divided into two cases: symmetric and asymmetric. When analyzing object similarity, the two state values of symmetrical binary attributes contribute to similarity analysis consistently, while the two state values of asymmetrical binary attributes have different weights thus their contributions to similarity analysis are inconsistent.

Table 1 shows all binary attributes of two objects. q means the number of binary attributes with 1 for object I and j; r is the number of binary attributes with 1 for object i and 0 for object j; s is the number of binary attributes with 0 for object i and 1 for object j; t is the number of binary attributes with 0 for object i and j. The total number of binary attributes owned by an object is p:

\[ p = q + r + s + t \]

If the objects i and j are described by symmetrical binary attributes, the distance between I and j is calculated by the following formula:

\[ d(i,j) = \frac{r + s}{q + r + s + t} \]

If both objects I and j are characterized by asymmetric binary attributes, and the weight is the highest when the attribute value is 1. Then t is considered negligible in the above publication, that is, the attribute value of 0 is subtracted from the total number of attributes, because these attributes have little significance in object similarity analysis. The distance calculation formula of I and j is as follows:

\[ d(i,j) = \frac{r + s}{q + r + s} \]

### 3. Parameter Learning of Reservation Registration

Parameter learning refers to the process of defining network parameters when the network topology is known. Various algorithms can realize parameter learning. The performance of learning algorithms used for samples is mainly judged according to their learning accuracy and speed. The binary attribute similarity method is mainly used for parameter learning of incomplete data sets and the algorithm is stable and simple to implement. Because the process of booking registration service is complex and involves many doctors, it is inevitable that some data will be missing in the process of booking registration service under the constraints of more or less factors, so that incomplete data sets will be formed after data observation. So this paper uses the binary attribute similarity method to parameter earning of the original data samples. Figure 1 shows the incomplete data processing flow.
Suppose that $D = \{D_1, D_2, \ldots, D_n\}$ denotes a set of missing sample data sets, $Z = \{Z_1, Z_2, \ldots, Z_n\}$ denotes hidden data sets; $X_i$ denotes missing samples in $D_i$; $\theta^i$ denotes the current estimation of parameter $\theta$, and defines $D_i$-based data sets $\theta$ as shown in the formula:

$$l(\theta|D) = \sum_{i=1}^{n} \sum_{x_i \in \partial D_i} P(X_i = x_i|D_i, \theta^i) \log P(D_i, X_i = x_i|\theta)$$

In this formula $P(X_i = x_i|D_i, \theta^i)$ represents when $X_i = \emptyset$, $P(X_i = x_i|D_i, \theta^i)$ equals to 1. In cyclic iteration, sample set $D$ is invariant, so $l(\theta|D, \theta^i)$ is often expressed by $Q(\theta|\theta^i)$.

4. Conclusion
This paper is based on the similarity of binary attributes which is used to solve the problem of hospital reservation registration service time. The relevant model is established, and the parameter learning method is elaborated. The data set is imported into the software platform for training and the learning results are obtained. The experimental results and model evaluation show that the proposed model can realize the dynamic estimation of reservation registration service time with high accuracy.

Acknowledgments
This paper was co-supported by the Foundation for Young Scholars in Wuhan Donghu University under grant 2018dhzk006. Thanks for the support of Wuhan Donghu University. Thank you for providing the experimental platform in Wuhan Donghu University. We have completed the implementation of the algorithm on this platform and the experimental data is presented.

References
[1] Yildirimoglu M, Geroliminis N. Experienced travel time prediction for congested freeways[J].
[2] Salamanis A, Meladianos P, Kehagias D, et al. Evaluating the effect of time series segmentation on STARIMA-based traffic prediction model[C] //Proc of the 18th IEEE International Conference on Intelligent Transportation Systems. Piscataway, NJ: IEEE Press, 2015: 2225-2230.

[3] Nath R P D, Lee H J, Chowdhury N K, et al. Modified K-means clustering for travel time prediction based on historical traffic data[C] //Proc of the 14th IEEE International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. Berlin: Springer, 2010: 511-521.

[4] Zhang Ting, Xia Yingjie, Zhu Qianqian, et al. Mining related information of traffic flows on lanes by K-medoids[C] //Proc of the 11th IEEE International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway, NJ: IEEE Press, 2014: 390-396.

[5] Lin Lei, Wang Qian, Sadek A W. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction[J]. Transportation Research, Part C: Emerging Technologies, 2015, 55(6): 444-459.

[6] Chen Ying, Kim J, Mahmassani H S. Pattern recognition using clustering algorithm for scenario definition in traffic simulation-based decision support systems[C] //Proc of the 17th IEEE International Conference on Intelligent Transportation Systems. Piscataway, NJ: IEEE Press, 2014: 798-803.

[7] Ma Xiaolei, Wu Y J, Wang Yinhai, et al. Mining smart card data for transit riders’ travel patterns[J]. Transportation Research, Part C Emerging Technologies, 2013, 36(11): 1-12.

[8] Mei Yu, Tang Keshuang, Li Keping. Real-time identification of probe vehicle trajectories in the mixed traffic corridor[J]. Transportation Research, Part C: Emerging Technologies, 2015, 57(8): 55-67.

[9] Ding Yi, Fu Xian. Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm[J]. Neurocomputing, 2015, 188(5): 233-238.

[10] Lyu Yinghua, Ma Tinghuai, Tang Meili, et al. An efficient and scalable density-based clustering algorithm for datasets with complex structures[J]. Neurocomputing, 2015, 171(1): 9-22.

[11] Huang He, Tang Qifeng, Liu Zhen. Adaptive correction forecasting approach for urban traffic flow based on fuzzy-mean clustering and advanced neural network[J]. Journal of Applied Mathematics, 2013, 2013(1): 1-7.