ANOMALY DETECTION IN SHIPMENT ITINERARY DATA

Baojun Qiu, John Yen
Intelligent Agents Laboratory
The Pennsylvania State University
bqiu@ist.psu.edu, jyen@ist.psu.edu

Keywords: Anomaly detection, Data mining, Markov model, Sequential data

ABSTRACT

Anomaly detection is important in many domains, e.g. fraud detection. Due to the difficult of precise definition and quantification of anomaly, anomaly detection is one of the most difficult tasks in data mining [1]. Significant effort has gone in anomaly detection in unstructured data (e.g. outlier detection) and graphic data (e.g. structural anomalies mining in graph-based data [2]). But detecting anomalous shipment itinerary is different because the itinerary data is sequential data. The sequential relationship between ports in an itinerary must be considered for anomaly detection. Besides, the type of carried cargo, travel time, the weather and the properties of corresponding ships (e.g. country, tonnage, ship type and etc), may also be crucial for detecting anomalous shipment itinerary. The sequential relationship and various affiliated information make it a complicated task to detect anomalous shipment itinerary.

Based on n-gram language model [3], we proposed a hybrid model to detect anomalous itinerary based on the sequential relationship between ports in itineraries as well as other affiliated information. Markov models are applied to handle sequential relationship and calculate the probabilities of itineraries to be anomalies. The choice of order of Markov model is important. A high order Markov models achieves a better accuracy but suffers from high computational complexity while a low order model has less computational cost but inaccurate. In order to tradeoff the accuracy and computation, backoff[3] method is applied to combine the high order and low order Markov models. Training data is very important for Markov model. But in most practical situations, the training data is limited. The lack of training data will cause underestimate of unobserved patterns. To solve this problem, smoothing methods are used to shave a little bit of probability mass from the frequent occurrences and pile it instead on zero counts. Other information like cargo and ship tonnage is also useful for improving the accuracy of anomaly detection. This information is integrated into Markov model as the conditions in calculating probabilities. In addition, time spent used in travel is also considered in our hybrid model by using the interpolation technique to combine it with Markov models. To decide the highest order of Markov models used in our hybrid model, we applied cross-validation and compared the perplexities for different order models.

Our experiments show our approach perform very well in practical shipment itinerary data.
REFERENCES

