Interactive Poster: Visual Mining of Business Process Data

Ming C. Hao, Daniel A. Keim*, Umeshwar Dayal, Joern Schneidewind
(ming.hao, umeshwar.dayal, joern.schneidewind)@hp.com
Hewlett Packard Research Laboratories

1. Motivation

Business process data is inherently large and complex, mast often too complex to be directly visualized. Usually the business operations consist of many steps and alternatives and every data instance may take a different path through the process. In Figure 1, we show a fraud analysis process schema. Note that this business process is a very simple one; realistic business processes are at least 10 times larger.

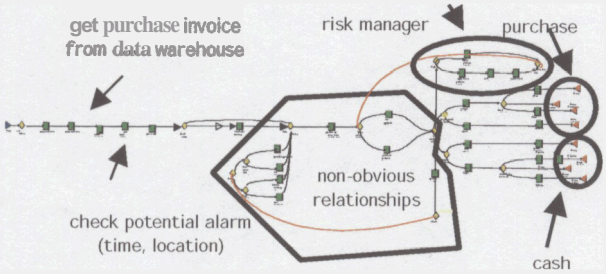


Figure 1: A Fraud Analysis Process

2. Our Approach

In this poster, we introduce a new interactive visualization technique to reduce data complexity by abstracting the most critical parameters, which influence business operations.

Our technique uses three basic components:

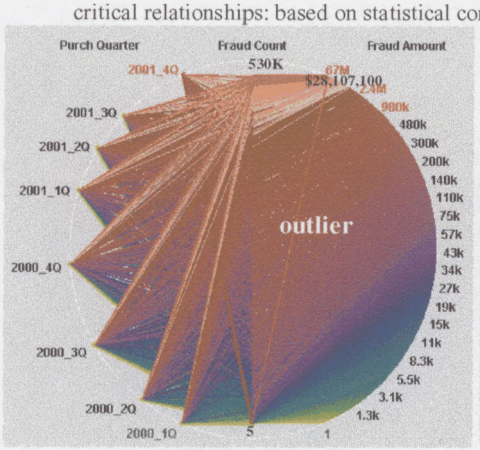
- a business parameter correlation matrix to determine the critical relationships: based on statistical correlation analyses, partial matching techniques, as well as cluster, and classification analysis.
- a triple-parameter symmetric circular graph layout to represent the source, intermediate, and destination attributes(nodes).
- a process flow matrix to link multiple circular graphs together to show the important portions of the business process operations.

In the credit card fraud analysis example shown in Figure 1, we may use a clustering algorithm to identify customer purchase quarter (2000_1Q to 2000_4Q), fraud count, issuer region, and sales type; and present them in relationships to the fraud amount as shown in Figure 2.

Our technique has been designed to work for large volumes of complex business process data. The automatic analysis determines the important relationships and the visualization shows the detected relationships. In contrast to Parallel Coordinate displays [I], we restrict the visualization to three parameters in one circular display and multiple displays are linked to show more than 3 parameters. In addition, we adapt a new layout to give more weight to important data values.

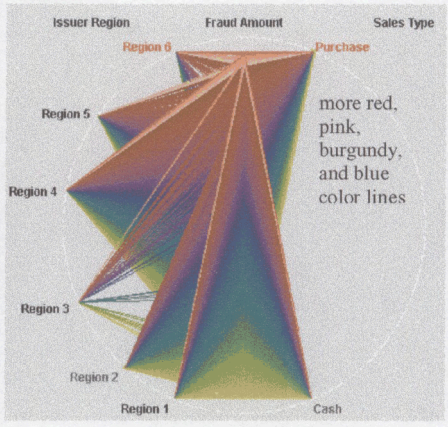
3. The Weighted Circular Graph Layout

In order to give important information more attention, we &fine a weight function for the nodes. This weight function w allocates more space for important nodes and is realized by a weighted radian *radian (i)*, as shown in Figure 3. The weight w_i of a node $v_i \in V_1$, $i \in \{1,..,N\}$, depends on a fourth attribute A. We &fine the weight by the ratio of v_i's attribute $a_i \in A$ and the sum of all attributes $a \in A$, $|A|=N$:



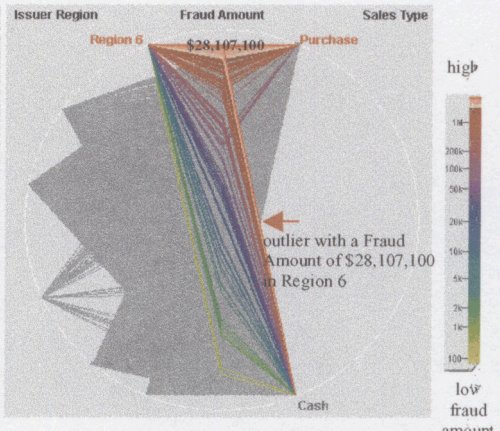
Step 1: Construct the first triple-parameter graph-Purchase Quarter, Fraud Count, and Fraud Amount.

- Shows fraud distribution in colored line patterns.
- Discovers an outlier - a red line crossing from low Fraud Count to high Fraud Amount (other lines are nearly parallel which corresponds to a high correlation) in 2001-4Q.



Step 2: Construct a 2nd triple-parameter graph-Issuer Region, Fraud Amount, and Sales Type.

- Region 6 (red) has the highest fraud amount (on the top of the circular graph, more pink and burgundy).
- More fraud comes from Purchase (more red, pink, and burgundy; less green) than Cash.



Step 3: This graph is generated when the analyst fades out all unrelated regions in step 2 for easy identification of the outlier. There is a red line crossing from Cash to the Fraud Amount of $28,107,100. This outlier is also linked to Fraud Count of 5 and Fraud Amount of $28,107,100 in Step 1. From this information, the analyst can learn that the cause of the outlier is a bad cash advance. This exceptional transaction might be a potential problem or error.

Figure 2: Fraud Business Operation Analysis

*Presently with the Computer Science Institute, University of Constance, Constance, Germany
keim@informatik.uni-konstanz.de

$$w_i = \frac{a_i}{\sum_{j=1}^{N} a_j}$$

$$radian(i) = \pi - \alpha\left[-\frac{1}{2} + w(i)\right]$$

$$w(i) = \sum_{j=1}^{j<i} w_j$$

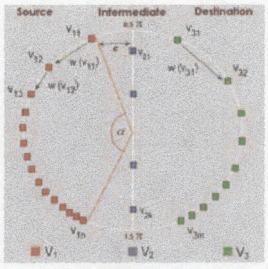


Figure 3: Computation of Weighted Node Positions

4. Applications

We have applied this approach to a number of real-world applications such as fraud analysis and IT service analysis

4.1 Fraud Analysis

To understand the cause of fraud, first, we select the three attributes with highest correlations from the business parameter correlation matrix, namely Purchase Quarter, Fraud Count, and Fraud Amount. Then, we select a 2nd set of highly correlated attributes. The analysis process is illustrated in Figure 2.

4.2 IT Service Analysis

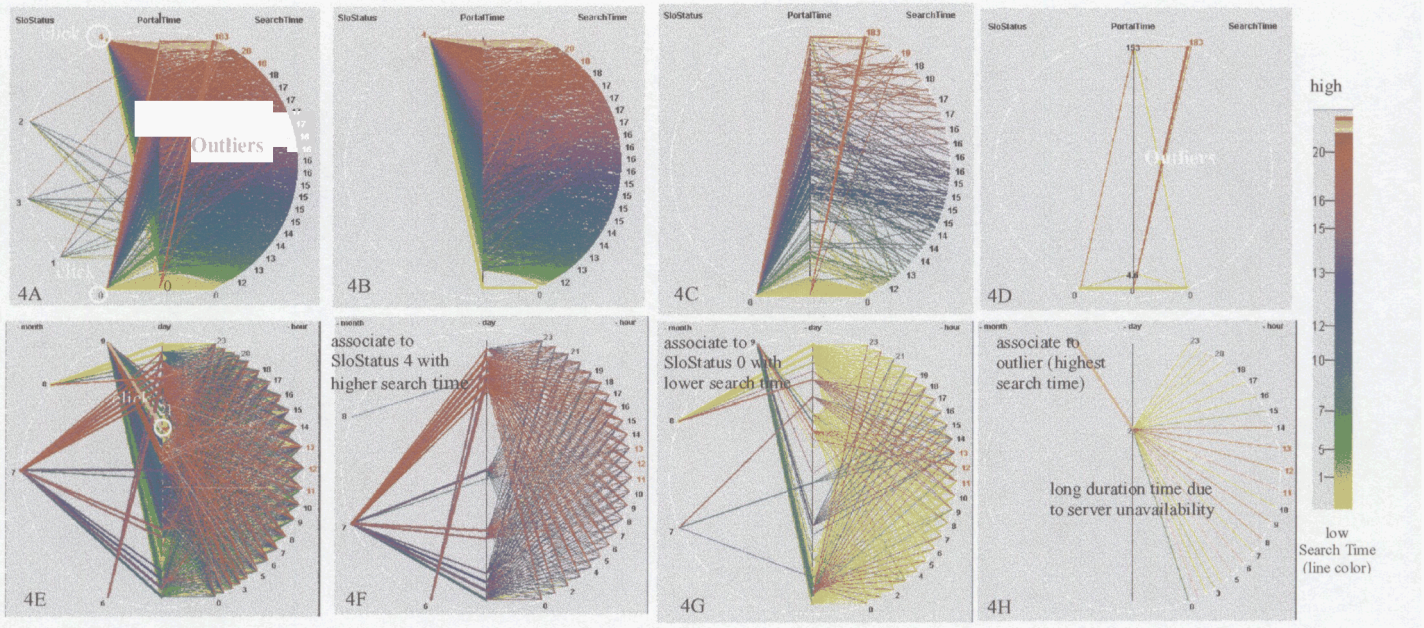
In IT service analysis, analysts are interested in the *cause* of unfulfilled services or the impact of failures on customer orders. Many research efforts have focused on how to transform the business process data, as logged by IT *services*, into valuable information. In this application, the SLO (Service Level Objectives) status indicates the probability of a service becoming unfulfilled (violated), *with* 0 being the most probable and 4 being the least probable. The data set contains 10,061 service transactions with over 50 SLO parameters. The most critical parameters are portal time, search time, month, day, and hour. Figure 4A-4H shows IT *search* time distribution, process flows, and the cause of the outliers.

5. Conclusion

In this poster, we reduce the complexity of business process analysis by abstracting the most critical parameters which influence business operations. We present them in multiple triple-parameter circular *graphs. Our* real world applications show significant advantages of our techniques in business process analysis.

Reference

[1] Inselberg A., Dimsdale B.: '*Parallel Coordinates:* a *Tool for Visualizing Multi-Dimemiom1 Geometry*', Visualization '90, San Francisco, CA, 1990.



- 4A-4D show a sequence of IT service process correlations

4A is generated from the three critical parameters (SloStatus, PortTime, and Search time) from the correlation matrix. PortalTime and SearchTime are highly correlated as seen by nearly parallel lines except *some* outliers crossing from low PortalTime to high SearchTime. Lints with the highest SearchTime and PortalTimes are colored red. SLOStatus 4 has the highest search time (more red, pink, burgundy) m 4A.

- 4E – 4H show the corresponding sequence of *search* time distributions (month, day, and hour)

4E is a second circular graph to show the time dependency. The processes in 4E-4H are linked to 4A-4D by the flow process matrix. When the analyst clicks *on* SLOStatus 4 in 4A, 4B and 4F are generated, showing that SloStatus 4 associates with higher search times, as seen by the blue and burgundy colors. When the analyst clicks on SLOStatus 0 in 4A, 4C and 4 0 are generated, showing that SloStatus 0 associates with lower search times, as seen by the yellow and green colors.

- 4D and 4H show the cause of outliers

4D and 4H are generated when the analyst clicks on day 21 (highest search time-more red lines) in Figure 4E. All unrelated lines are faded out. The *analyst* moves the pointer *on* the red lines end nodes (connected from day 21 to month 9, and hour 11-14) to drill down to the transaction record level information, such as server availability.

Figure 4 IT Service Process Visualization