First publ. in: Information visualization 5 (2006), pp. 15-27

Business process impact visualization and anomaly detection

Ming C. Hao[1]
Daniel A. Keim[2]
Umeshwar Dayal[1]
Jörn Schneidewind[2]

[1]Hewlett Packard Labs, Palo Alto, CA, U.S.A;
[2]University of Konstanz, Konstanz, Germany

Correspondence: Daniel A. Keim, Computer Science Institute, Universität Konstanz Box D78, Universitätsstr. 10, 78457 Konstanz, Germany.
Tel: +49 7531 88 3161;
Fax: +49 7531 88 3062;
E-mail: keim@inf.uni-konstanz.de

Abstract
Business operations involve many factors and relationships and are modeled as complex business process workflows. The execution of these business processes generates vast volumes of complex data. The operational data are instances of the process flow, taking different paths through the process. The goal is to use the complex information to analyze and improve operations and to optimize the process flow. In this paper, we introduce a new visualization technique, called *VisImpact* that turns raw operational business data into valuable information. *VisImpact* reduces data complexity by analyzing operational data and abstracting the most critical factors, called impact factors, which influence business operations. The analysis may identify single nodes of the business flow graph as important factors but it may also determine aggregations of nodes to be important. Moreover, the analysis may find that single nodes have certain data values associated with them which have an influence on some business metrics or resource usage parameters. The impact factors are presented as nodes in a symmetric circular graph, providing insight into core business operations and relationships. A cause–effect mechanism is built in to determine 'good' and 'bad' operational behavior and to take action accordingly. We have applied *VisImpact* to real-world applications, fraud analysis and service contract analysis, to show the power of *VisImpact* for finding relationships among the most important impact factors and for immediate identification of anomalies. The *VisImpact* system provides a highly interactive interface including drilldown capabilities down to transaction levels to allow multilevel views of business dynamics.
Information Visualization (2006) **5,** 15–27. doi:10.1057/palgrave.ivs.9500115

Keywords: Business intelligence; visual data mining; visualization technique

Introduction

What is the cause of an unfulfilled contract? If the problem continues for a day, how much will it cost? Which customer orders will be impacted? To answer such questions, many research efforts have focused on how to transform the business process data, as logged by the IT services, into valuable information. However, owing to multiple factors and the complexity of business operations, analysts face the challenge of understanding the underlying data and finding the important relationships from which to draw conclusions.

Charts, scatter plots, and spreadsheets are commonly used, but are inappropriate to visualize relations between multiple business parameters in a single view. Thus, analysts usually need to view pages of charts and reports to obtain the information they need to make informed decisions. However, the degree of difficulty increases as business operation complexity grows. Therefore, analysts need tools that provide them insights into

important business factors in the context of their overall business operations, for example, business operations on which these business factors depend on, or the impact of the changes of business factors.

In this context, visualization technology plays an important role and a number of advanced visualization techniques for business impact analysis have been proposed in the past. Parallel coordinates,[1] for example, is a geometric projection technique used for multidimensional visualization and automatic classification. SeeSoft's line representation technique[2] is used to visualize Year 2000 program changes. ILOG JViews is a tool for analyzing workflow processes. E_Bizinsights is used to provide dimension-based structures for web path analysis. All these example techniques can effectively display data for business decision making. The techniques start with the presentation of whole business operation and provide drill down capabilities to the analyst to obtain detailed information.

In this paper, we introduce a new visualization technique called *VisImpact*. This technique extends existing approaches by providing an integrated analysis of business process schema and business process instances in order to improve business operations. First, *VisImpact* analyzes the business process schema and data to identify the important factors that influence the business metrics and resource parameters. These important impact factors and the corresponding business process instances are then represented as nodes and lines on a symmetric circular graph to display the important relationships and details of the process flows. *VisImpact* is different from ILOG JViews workflow visualizations in that JViews displays the process workflow schema as a standard graph and uses standard charts to plot highly aggregated summaries of the workflow execution data. Our approach in contrast uses advanced analysis techniques to determine the important impact factors, which are then visualized as circular graphs. After presenting the basic idea of the *VisImpact* technique in the next section, we introduce the analysis techniques integrated in *VisImpact* for extracting relevant factors from business processes (Section on Determining the impact relationships) and describe how these information is mapped to circular graphs (Section on The circular VisImpact graph). In the Section on The VisImpact system we describe the *VisImpact* system and show its applications to real-word fraud analysis and business contract analysis with promising results (Section on Applications). Section on Evaluation and comparison evaluates the proposed system. A case study showing additional applications of the technique is presented in Hao *et al*.[3]

Basic idea of VisImpact

Business operation data is inherently complex, most 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. To understand the business process

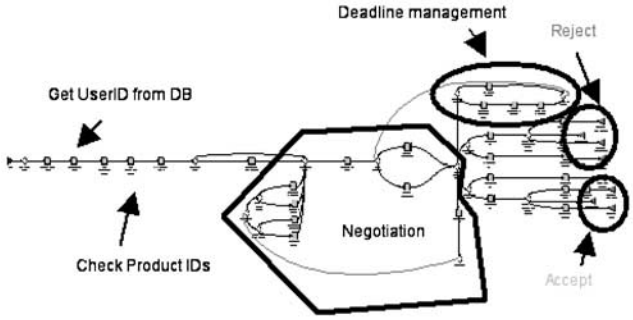

Figure 1 A product order activity workflow.

and the impact of certain parameters, in the first step *VisImpact* analyzes the process data. As an introductory example, Figure 1 shows a simplified version of a product order activity process schema. Note that this business process is a very simple one; realistic business processes are at least 10 times larger. To simplify the complexity of the visualization, *VisImpact* automatically abstracts important attributes, called impact factors, using data mining techniques ranging from statistical correlation analysis and partial matching to techniques used in clustering and classification analysis. Additionally, the analyst may steer the analysis process by interactively selecting business attributes and metrics of interest manually, and the system detects impact factors for the corresponding selection. The analysis results are then stored in an impact factor matrix. Finally, *VisImpact* transforms these impact factors to nodes, with lines between nodes on a symmetric circular graph, representing a portion of the business operation.

The goal was to provide an appropriate visual layout, that is able to represent an abstraction of the underlying complex workflow, but at the same time shows the relationships between discovered relevant business impact factors for further visual analysis. In general, each business workflow can be modeled as directed graph containing three different categories of nodes: start nodes, end nodes, and inner nodes including different node types like decision or action nodes. The edges between the nodes define the process flow. Typically the process flow starts at a start node, passes certain inner nodes and ends at an end node. We adapted this principle by generating circular graph layouts consisting of three types of nodes, representing three types of nodes of the underlying business process schema. The edges between the nodes represent the process instances and their color represents an additional attribute, for example, a business metric. Thus, each circular graph visualizes four different attributes:

- *Source Attribute* for partitioning the left side of the circle;
- *Intermediate Attribute* for partitioning the center axis;
- *Destination Attribute* for partitioning the right side;

- *Color Attribute* (colored lines) for visualizing business metrics, such as response time, dollar amount, and contract fulfillment degree.

Each process instance is shown as a line-connecting source, intermediate and destination nodes. *VisImpact* records the process instances in the process flow map, so that analysts are able to track the cause–effect paths in real time. Each circular graph represents a special impact relationship, and multiple circular graphs are linked together to show whole business process operations.

In the product order activity example, *VisImpact* uses a clustering algorithm to identify types of customers (Gold, Silver, Regular) and order processing duration times (very long, long, medium, short) and present it in relationship to the outcome of the decision (Accept, Reject, Cancel) as shown in Figure 2.

The lines are colored according to the duration time. In the example, it is obvious that the higher the rank of the customer, the faster the processing of the order and the higher the likelihood of an acceptance decision. Since the data is shown on an instance level (and not in an aggregated form) it becomes clear that for a regular customer the duration time varies largely and that the likelihood of a reject decision increases with higher processing durations. Note that there are also some exceptions to this general tendency which can be seen by the red line that ends in the accept node. The second *VisImpact* graph (see Figure 3) shows a second impact relationship, namely between specific instances of the negotiation nodes, duration time, and outcome of the decision. Note that in this case the negotiators are partitioned into two specific negotiators A_1 and A_2, then a group of negotiators AG, and all other negotiators *Others*. The four groups have been identified by the automatic algorithm based on the similarity of the business process data instances.

In the third graph (see Figure 4), we show how *VisImpact* can be used to analyze product order contract fulfillment, that is, the ability to deliver certain products within a specified time period. There is a penalty cost if a contract is delayed beyond a certain number of days. The *VisImpact* system determines that penalty cost is closely

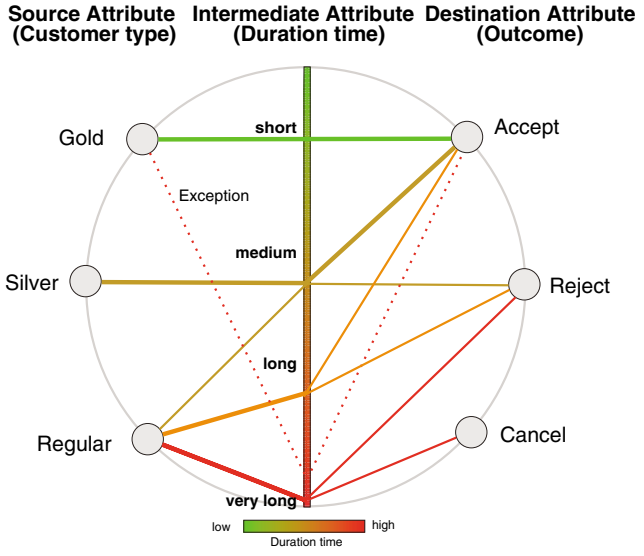

Figure 2 Product order activity by process duration times (color represents process duration time).

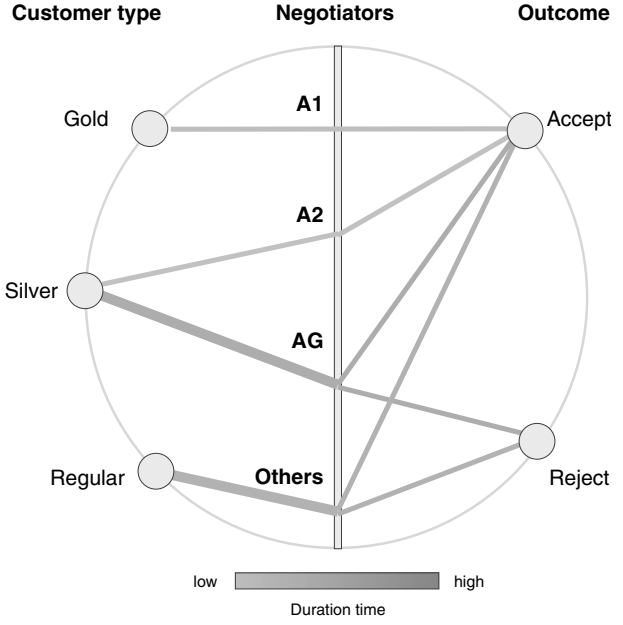

Figure 3 Order activity by negotiators.

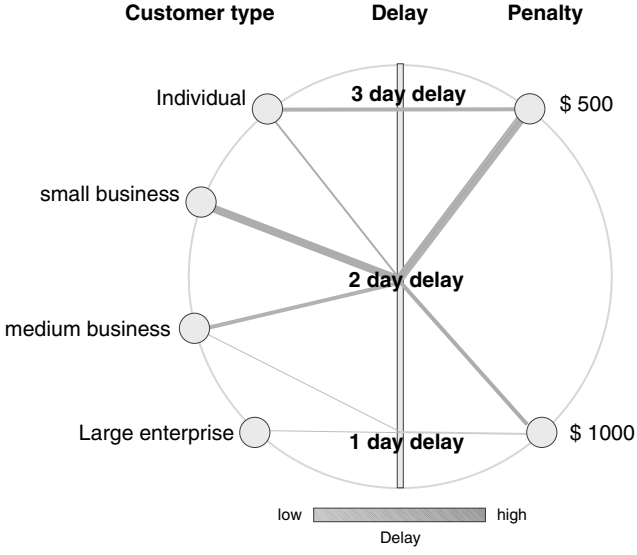

Figure 4 Order activity by delay.

related to specific delay times. Note that for correlating penalty cost and delay time to customers, the system proposes a different classification of the customer nodes, namely *large enterprise, medium business, small business,* and *individual*. To make large volumes of data easy to explore and interpret, *VisImpact* links multiple circular graphs and records the path of process instances in the process flow map. Therefore, analysts can quickly click on a node or a line to observe all linked operation flows across different graphs in real time.

Formal definition of VisImpact

In the following, we formally introduce the techniques used to generate *VisImpact* visualizations. The first step is the identification of all relevant impact relationship. For this step we perform a global correlation analysis and use partial matching, cluster and classification analysis techniques. The result of this step are triples of related attributes which are then visualized as nodes and the instances are represented as edges of a graph. The problem is to find a good graph layout that supports human problem solving and decision-making processes. There are some general requirements that graph layouts for human consumption should fulfil, known as aesthetics criteria.[4] Some important criteria for *VisImpact* are display symmetry, edge crossing reduction, uniform vertex distribution, and uniform edge lengths. Additionally, the layout should present an ordering of the nodes corresponding to the business parameters and present an abstraction of the data. In addition, the layout should allow a visualization of large data volumes. A circular layout is chosen, because it provides a good compromise between all requirements.[5]

Determining the impact relationships

The first step of *VisImpact* is to determine the most important impact relationships. For this step we use (semi-) automatic data mining techniques, namely statistical correlation analysis, partial matching techniques, as well as cluster and classification analysis.

Statistical correlation and similarity analysis First, we determine the pairwise global correlations among all measurements as given by Pearson's correlation matrix. Pearson's correlation coefficient r between bivariate data, A_{1i} and A_{2i} values ($i = 1, \ldots, n$) is defined as

$$r = \frac{\sum_{i=1}^{n} (A_{1i} - \bar{A}_1)(A_{2i} - \bar{A}_2)}{\sqrt{\sum_{i=1}^{n} (A_{1i} - \bar{A}_1)^2 \sum_{i=1}^{n} (A_{2i} - \bar{A}_2)^2}}, \qquad (1)$$

where $\bar{A}_1$ and $\bar{A}_2$ are the means of the A_{1i} and A_{2i} values, respectively. If two dimensions are perfectly correlated, the correlation coefficient is 1, in case of an inverse correlation -1. In case of a perfect correlation, we can omit one of the attributes since it contains redundant information. In most cases, however, the correlations are not perfect and we are interested in high correlation coefficients and select sets of three highly correlated attributes to be visualized in *VisImpact*. Other statistical correlation coefficients such as the Spearman's correlation are provided in *VisImpact* as well.

An available alternative for adjacently depicting similar dimensions is to use the normalized Euclidean distance as a measure for global similarity Sim_{global} defined as

$$Sim_{Global}\ (A_i,\ A_j) = \sqrt{\sum_{i=0}^{N-1} (b_i^1 - b_i^2)^2}, \qquad (2)$$

where

$$b_i^j = \frac{a_i^j - MIN(A_j)}{MAX(A_j) - MIN(A_j)}.$$

In order to become more robust against outliers, instead of using MAX (the 100%-quantile) and MIN (the 0%-quantile), we use the 98 and 2% quantile of the attribute. The global similarity measure compares two whole dimensions such that any change in one of the dimensions has an influence on the resulting similarity. The defined similarity measure allows it to determine triples of similar attributes for the successional visualization. Since in general, computing similarity measures is a non-trivial task, because similarity can be defined in various ways and for specific domains, the modular design of the *VisImpact* system allows the integration of specific similarity measures with little effort, like similarity measures proposed in the context of time series data[6] or similarity measures presented in.[7]

Partial similarity In real business process applications global similarities are rare, since in most cases correlations only occur for certain subsets of the data. Imagine for example two business measures over time, like the duration time for Gold and Silver customers in Figure 2. There may be short periods where the two measures show a similar behavior, for example, because of some global development. However, it is unlikely that they behave similar over days or weeks. In the impact relationship analysis we therefore have to analyze the data for partial similarities. In our application scenario, we are especially interested in periods where two attributes behaved similar. Thus, given the two variables A_k and A_l, the synchronized partial similarity[8] measure is employed , to detect pairwise attributes with periods of similarity in the data:

$$Sim_{Sync}(A_k, A_l) = \max_{i,j} \left\{ (j-i) | (0i<j<N) \sqrt{\sum_{z=i}^{j} C_z < \varepsilon} \right\}, \qquad (3)$$

where $C_z = (b_z^k - b_z^l)^2$, with b_i^j defined as above and $\in$ is some maximum allowed dissimilarity. This partial similarity measure uses the length of the longest sequence that is at least $\in$-similar (under scaling and translation invariance). Triples of attributes with pairwise maximum Sim_{Sync} values are then selected for *VisImpact* analysis. Depending on the application, the partial similarity may

also be an Unsynchronized Partial Similarity.[8] In this case, two dimensions do not have to be similar at the same 'time but in an arbitrary time frame of the same length. Since computing partial matchings is a time-consuming process, most approaches like those proposed by Yang *et al.*[9] and Faloutsos *et al.*[10] also use some heuristics and index structures to speed up the computation,[11] that will be considered in future extensions of *VisImpact*.

Cluster analysis For some attributes, the parameter values are continuous (such as dollar amount), for others, there are large numbers of categorical values (such as expense requestors). In order to perform a useful impact analysis, it is important to partition the value ranges appropriately. Cluster analysis can help to do this based on the characteristics of the data instances. The cluster analysis may, for example, find out that – based on the characteristics of their product order flows – the companies may be partitioned into three groups (gold, silver, regular) and the negotiators into two single ones (A1, A2) and two groups (AG, Others).

There are a large number of clustering methods which have been proposed in the literature, one of the most general techniques is kernel density estimation.[12] In kernel density estimation, the influence of each data point is modeled using a kernel function, and the overall density of the data is calculated as the sum of the kernel functions of all data points. Clusters can be derived from a density function by density based single linkage or hierarchical clustering. Owing to the large number of analyses that need to be performed in the *VisImpact* framework, we have to use an efficient implementation of kernel density estimation, and therefore the DENCLUE algorithm[13] is employed.

Classification analysis In some applications, the goal of the data exploration is to understand the relationship between the business process data and some specific business metrics such as response time, dollar amount, or degree of contract fulfillment. If the analyst is interested in a specific business metric, we can perform the automatic analysis with the business metric as target attribute. The task is to find the business process parameters that are best predicting the outcome of the target attribute. A well-known heuristic for this task is the GINI index, which is also used in decision-tree construction. Given a business metric B which is partitioned into a disjoint set of k classes (e.g. accept, reject) or value ranges (e.g. large, medium, small) denoted by $C1, \ldots, C_k$, $(B = \cup_{i=1}^{k} C_i)$, then the GINI index of an attribute A which induces a partitioning of A into $A1, \ldots, A_m$ is defined as

$$InfoGain_{GINI}(B, A) = \sum_{i=1}^{m} \frac{|A_i|}{|B|} GINI(A_i), \qquad (4)$$

where

$$GINI(A_i) = 1 - \sum_{j=1}^{k} \left[\frac{|C_j|}{|A_i|}\right]^2.$$

The *InfoGain* is determined for all attributes and attribute combinations and the two attributes with the highest *InfoGain* with respect to the target attribute B are chosen for visualization. Alternatively, we use the attribute A_x with the highest *InfoGain* and then repeat the calculation with A_x as target attribute to find the second attribute to be displayed.

The circular VisImpact graph

The business impact visualization is defined as a graph $G = (V, E)$, where V is a set of nodes connected by edges E. The node set V is partitioned in k subsets $V_1, \ldots, V_k$ depending on k partitioning attributes. Each edge $(u, v) \in E$ implies either $u \in V_i$ and $v \in V_{i+1}$ or $u \in V_{i+1}$ and $v \in V_i$, $i \in 1, \ldots, k$–1. The nodes V represent the set of data items for the corresponding k classes of V and the edges represent the relationships and interactions between them. An edge can have at least two attributes, showing characteristics of the relationship, represented by width and color of the edge.

In the *VisImpact* System, a special case of circular graph is used, where the node set V of the graph consists of three subsets V_1, V_2, V_3, $V = V_1 \bigcup V_2 \bigcup V_3$, $(V_i \bigcap V_j = \phi \Rightarrow i \neq j)$. The set of source nodes V_1, is determined by the first attribute (source attribute). The second attribute (intermediate attribute) determines the subset V_2 of intermediate nodes, and the third attribute (destination attribute) determines V_3, the set of destination nodes. Corresponding to the definition of the general circular graph, there exist only edges $e = (u, v) \in E$ between V_1 and V_2 or V_2 and V_3. In order to present the given nodes and edges in a circular layout, let $C = (x, y, r)$ be a circle with center (x, y) and radius r in the $2D$-plane. We introduce a screen positioning function $f: V \rightarrow R^2$, which determines for each node $v \in V$ the x/y-position $(v.x, v.y)$ on the circle.

Since we want to visualize the relations and interactions between three sets of nodes, we divide the circle C in three regions to place the nodes from the three sets V_1, V_2, V_3. The nodes of V_1 are placed on the left side and the nodes of V_3 are placed on the right side of the circle, which means for all nodes $v \in V_1 \bigcup V_3$ holds:

$$Cr^2 = (v_x - C_x)^2 + (v_y - C_y)^2. \qquad (5)$$

For all nodes $v_i \in V_1$ is $v_i.x – Cx < Cx$ and for all nodes $vj \in V_3$ is $v_j.x – Cx > Cx$. The nodes of V_2 are placed on the center axis of the circle, which means on a line from Point $P1(Cx, Cy–Cr)$ to Point $P2(Cx, Cy + Cr)$, so that for all $v_j \in V_2$

$$v_j x = Cx \quad \text{and} \quad Cy - Cr < v_j y < Cy + Cr.$$

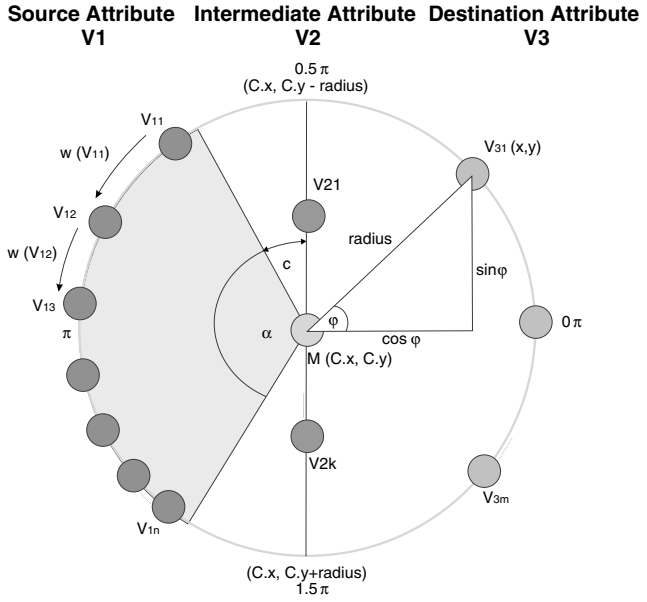

Figure 5 Computation of node positions on the circular *Vismap* layout.

Computing the node positions The placement of nodes on the circle axis is straightforward and depends only on the selected mapping. To place nodes on the left or right half of the circle, the positioning function f employs the radian ϕ to compute the position for each node depending on the selected mapping, as shown in Figure 5. For quantitative data, linear mapping is used to map the data points to the left side, the right side or the center axis of the circle, and the radian is determined accordingly. Optional, the data points can be placed in an ordered equidistant manner. This is especially useful for categorical data or in cases where the analyst is more interested in the process flow than in exact node values. The radian ϕ for a node $v_i \in V_1$ is then defined as follows:

$$\phi = \pi - \alpha\left[-\frac{1}{2} + \frac{i}{n}\right], \quad i = 0, \ldots, n. \tag{6}$$

The angle α, $0 < \alpha < \pi$, describes the positioning area of the nodes, shown in Figure 5. In order to position the nodes of V_1 on the left side of the circle (i.e. $0.5\pi < \phi < 1.5\pi$), we set $\alpha = c\pi$, $0 < c < 1$. For placements on the right side of the circle, that is, for positioning of all nodes $v \in V_3$, π–α has to be replaced by α in the equation above. The parameter c separates the nodes on the right and left half of the circle and the nodes in the middle. The term i/n divides the drawing area, given by α, in n equidistant locations in order to place the n nodes from V_1. The radian ϕ is used to compute a position for each node $v_i \in V_1$:

$$v_i x = Cx + \cos(\phi) \cdot Cr,$$
$$v_i y = Cy + \sin(\phi) \cdot Cr.$$

The node positions for the nodes $v \in V_3$ can be computed analogical. Color coding and tool tip techniques are used to represent relevant node attributes.

Weighted node positions In order to give important information in our visualization more attention, an optional weight function can be used. Instead of just ordering the nodes according to their values and then place the nodes on the circle in an equidistant manner, this weight function gives important nodes more space on the screen while less important nodes get less space, realized by a weighted computation of the radian ϕ, as shown in Figure 5. The weight $weight_i$ of a node $v_i \in V_l, i \in (1, \ldots, N)$, depends on a forth attribute A. We define the weight by the ratio of v_i's attribute $a_i \in A$ and the sum of all attributes $a_j \in A$, $|A| = N$, where $i \in (1, \ldots, N)$:

$$weight_i = \frac{a_i}{\sum_{j=1}^{N} a_j}. \tag{7}$$

After computing a weight for each node, *VisImpact* orders the nodes by their weights and places them by starting at the top of the circle. The weighted positioning function w distributes the available space on the circle to the nodes by calculating a weighted radian ϕ_{weight} for each node $v_i \in V_1$, $i \in (1, \ldots, N)$:

$$\phi_{weight} = \pi - \alpha\left[-\frac{1}{2} + w(i)\right],$$
$$w(i) = \sum_{j=0}^{j<i} weight_j. \tag{8}$$

In order to place the nodes in V_3 on the right side of the circle, π–α has to be replaced by α in the formula above.

Placement of categorical attributes In cases where the ordering of nodes in the *VisImpact* visualization is not implicitly given by the node values, for example, for categorical attributes like *customer name* or *customer type,* the analyst is typically only interested in the process flow between certain attributes. The goal then is to find a circular node layout that reduces edge crossings, since they may reduce the readability of the resulting graph. Therefore, a placing method that reduces edge crossings by rearranging single nodes is integrated into the *VisImpact* system to place nodes with no implicit ordering. Since in general, the problem of finding vertex orderings that minimize edge crossings in a layered graph is NP-hard, even for three-layered graphs as used by *VisImpact,*[14] heuristics are needed to solve even moderately sized problems.

Let $G = (V, E)$, $V = V_1 \bigcup \ldots \bigcup V_k$, $V_i \bigcap V_j = \phi \Leftrightarrow i \neq j$, be a general circular graph as described above. An ordering layer V_i, $i \in (1, \ldots, k\text{–}1)$ is specified by a permutation π_i of *Vi.* We express the ordering of V_i by the permutation π_i. Let *cross* $(G, \pi_1, \ldots, \pi_k)$ be the number of edge crossings in

a straight line drawing of G given by $\pi_1, \ldots, \pi_k$. The minimum number of edge crossings that can be achieved by reordering the vertices in $V_1, \ldots, V_k$ is denoted by $opt(G)$:

$$Opt(G) = \min_{\pi_1,\ldots,\pi_k} cross(G, \pi_1, \ldots, \pi_k). \quad (9)$$

Having three sets of nodes V_1, V_2, V_3, *VisImpact* computes a minimal edge crossing by dividing this three-layered crossing minimization problem into two two-layered One Sided Crossing Minimization Problem:

$$Opt'(G) = \min_{\pi_i,\pi_i+1} cross(G, \pi_i, \pi_i+1), \quad i = 1, 2. \quad (10)$$

$Opt'(G, \pi_i)$ denotes the minimal attainable number of edge crossings by fixing the permutation of V_i and reordering the nodes of V_{i+1}. The Barycenter heuristic[15] is used to compute such a node ordering. The basic idea of this heuristic is to simply compute the average position, that is, the Barycenter, for each node and then sort the nodes according to these numbers. In typical application scenarios not all three attributes will be nominal or categorical without given orders, which restricts the crossing minimization process.

The VisImpact system

System architecture and components

To analyze large volumes of transaction data with many impact factors, *VisImpact* has been integrated into the visual data mining system *VisMine*.[16] The system uses a web browser with a Java activator to allow real-time interactive visual data mining over the web. The *VisImpact* architecture contains four basic components:

1. Abstract component
   The abstract component of *VisImpact* derives an *impact factor matrix* from the input data. The system can then automatically show a number of VisImpact visualizations based attributes with high-impact relationships. Alternatively, the analysts can select a pair of impact factors from the matrix based on the knowledge of the user and the application requirements. Then, from the pair of impact factors, *VisImpact* automatically abstracts the third impact factor that has the highest impact value with the selected pair of impact factors. In addition, *VisImpact* provides a *control window* to allow analysts to real-time select impact factors for further analysis.
2. Layout component
   This component orders, groups, maps, and weights the abstracted impact factors and transforms their relationships to lines between two nodes according to the *process flow map*. Nodes and lines are laid out on a symmetric circular graph. The width of the line represents the number of lines with the same process flow. The color of a line is the average value of a data item. Nodes and lines on the circular graph are weighted by the average value of a data item. Nodes with higher weights are given more space on the graph. The label of the nodes with the highest weights is colored red. Lines with the highest weights are drawn last to avoid overlapping.
3. Interaction component
   To make the circular graph easy to explore and interpret, *VisImpact* provides fade-in, fade-out, clicking, and drill-down capabilities.
4. Extension component
   Additional circular graphs are generated as the number of selected attributes grows.

Anomaly detection

VisImpact employs a *process flow map* to link related process flows and relationships across different circular graphs. *VisImpact* instantly detects exceptions by finding red lines (exceed threshold) or crossed lines (anomaly) in a graph. For example, in fraud analysis (Section on Fraud analysis), a high fraud amount usually is associated with a high fraud count. There could be a potential problem if a high fraud amount occurs with a low fraud count as shown later in Figure 6a. *VisImpact* provides the following visual capabilities:

- Fade-in and fade-out
  VisImpact allows analysts to focus on an outlier and fade in related process paths. The unrelated nodes and flows are faded out. The analyst can easily discover the source of the problem by tracing the lines starting from the anomaly.
- Drill down
  The analysts select a single node or a single line to drill down to the transaction level to display multilevel views of business dynamics.

Applications

We have experimented with *VisImpact* for fraud analysis, a case study is presented in,[3] and service contract analysis using real-word business data.

Fraud analysis

Fraud is one of the major problems faced by many companies in the banking, insurance, and telephony industries. Over $ 2 billion in fraudulent transactions are processed yearly on electronic payments. Transforming raw transaction data into valuable business operation information to enable fraud analysis will save companies millions of dollars. Fraud analysis specialists require tools that help them to better understand fraud behavior and impact factors as well as to identify unusual exceptions. Typical questions in fraud analysis are:

1. What is the fraud growth rate in recent years and what are the impact factors?
2. Which sales region and sales type has the most fraud?
3. Are there any outliers and what is their cause–effect?

To address these three questions, *VisImpact* first selects three highly correlated attributes from the impact factor matrix such as purchase quarter, fraud amount (aggregated), and fraud count. Using these three impact factors, *VisImpact* lays out the nodes and flows in a circular graph as shown in Figure 6a. Figure 6a shows that there are high correlations (more parallel lines) between fraud amount and fraud count. Colors represent the value of the fraud amount; red represents a fraud amount that is in the top 10%. Most important, there is an outlier (a red line) crossing from low fraud count (5 counts) to a very high cash advance with a fraud amount of $ 28,107,100. This exceptional transaction might be a potential problem or error.

To understand which sales regions or sales have the most fraud, the analyst selects the region as the source node and sales type as the destination node from the user domain knowledge and draws a second circular graph as shown in Figure 6b. From Figure 6b, it can be learned that region 6 has the highest fraud amount with more red, pink, burgundy and blue lines than other regions. Purchase has a higher fraud amount than cash advance. This is because purchase has many more red and burgundy lines than cash advance.

To find outliers and their root-cause, *VisImpact* uses the process flow map to identify related operation paths of a transaction record and to discover exceptions. Most interestingly, an outlier is seen as a red line crossing from cash to the high fraud amount. The analyst can easily move the pointer to find detailed information about this outlier, such as the amount and purchase quarter. Investigating further, the analyst can select the region 6 node to focus only on region 6 fraud. The fraud from other regions is faded out as shown in Figure 7. The analyst can quickly see that the outlier comes from cash advance. This capability to trace the process flow of a transaction is crucial for finding the cause–effect relationship of outliers. Using the above information, the company is able to place strict control on certain regions

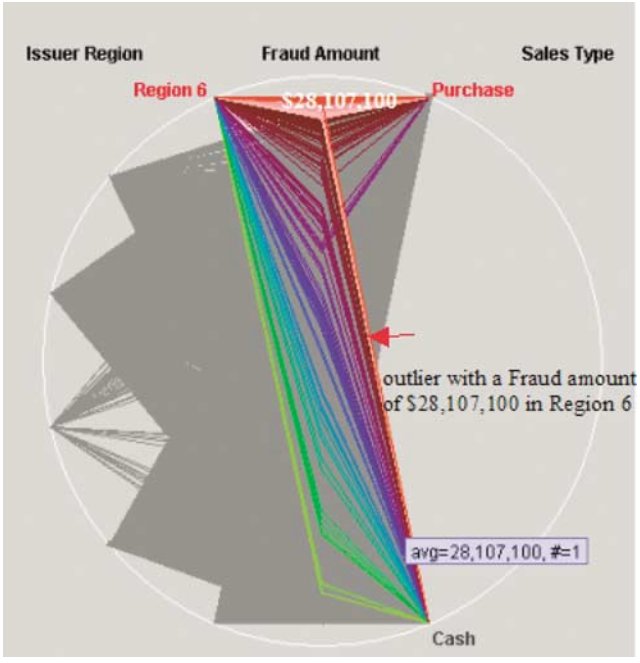

Figure 7 *Finding the cause of the outlier:* Region 6 from Figure 6b is selected. The outlier is seen as a red line crossing from cash to the amount of $ 28,107,100. The outlier is linked to region 6 from the left half of the graph. The outlier is also linked to fraud count 5 and 2000-4Q in Figure 6a. The cause of the outlier is a cash advance that happened in 2000-4Q, region 6 with a fraud amount of $ 28,107,100 and fraud count of 5.

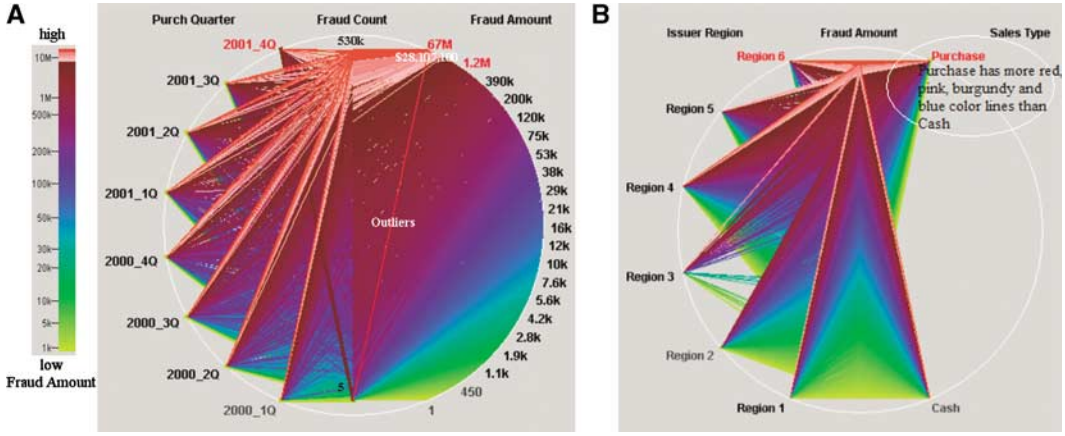

Figure 6 Fraud analysis. (A) *Fraud distribution by purchase quarter:* Impact factors are quarter, fraud amount and fraud count. Each line is a transaction, color represents the value of the fraud amount: fraud increases with time (e.g. more red lines in each quarter). 2001-4Q has the highest fraud amount (red label). An outlier: a red line crosses from low fraud count to high fraud amount (other lines are nearly parallel – high correlation) (B) *Fraud distribution by region:* Using fraud amount from (A) and choose two other impact factors: region and sales type. Region 6 (red) has the highest fraud amount (on the top of the circular graph, more red, pink, and burgundy; less green lines). Most fraud comes from purchase *vs* cash. Purchase has more red, burgundy, and blue lines than cash.

(countries) and credit card usages. After better understanding the source of the fraud, the company will be able to take preventive action.

Service contract analysis

All businesses have relationships with customers and suppliers; they execute business processes to obtain services from suppliers and add value to deliver services to customers. Such service processes are usually modeled by service level objectives (SLOs) and contracts,[17] stipulated between customers and suppliers. A contract typically contains SLOs defining what service should be delivered with what level of quality and within what time period. An important question business managers need to pursue is whether their business operations are fulfilling the SLOs. This is a difficult problem, often complicated by service performance (e.g. response time, server availability) involved in the execution of business operations.

We have applied *VisImpact* to a real-world, large-scale data set of service contract analysis in an effort to better understand SLO operational flows, distributions, and anomalies. In this application, the SLO status indicates the probability of a contract 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 impact factors such as SLO status, portal response time, search response, month, day, and hour. *VisImpact* maps nodes to SLO impact factors, lines to service transactions, line widths to the number of service transactions, and colors to the values of selected impact factors (i.e., search response time). Nodes are placed in order according to the selected impact factors.

Operational flows and their distribution *VisImpact* first abstracts the three most highly correlated factors from the impact factor matrix and constructs a circular graph as shown in Figure 8A. The source nodes show *SLO status,* the intermediate nodes *portal response time,* and the destination nodes *search response time.* Nodes are connected with lines from the process flow map. The color of a line is the value of the *search response time* (ms). Figure 8a shows that the SLO status is highly impacted by both portal and search response times. A transaction with high values for search response time often has high values for portal response time as shown by the numerous nearly parallel lines. Note that there are major outliers, shown by the red lines crossing from a low portal response time to a high search response time.

Time dependency *VisImpact* generates a second circular graph to show the time dependency as presented in Figure 8b from user domain knowledge. The transaction process flows in Figure 8b are tightly linked to SLO status in Figure 8a. In Figure 8b, the source nodes are month (6,7,8,9), the intermediate nodes are days (1–31), and the destination nodes are hours (0–23). Figure 8B shows search response time distribution over time. The color denotes the search response time.

Process flow relationships between multiple circular graphs *VisImpact* helps to discover that the search response time is the potential root-cause of the unfulfilled SLOs. This is seen through the linking of process flows across two circular graphs. As shown in Figures 9a–d, we are able to verify this relationship because (a) in Figure 9a and b the higher probability SLO status (e.g., SLO status 4) is associated with slower response times, as

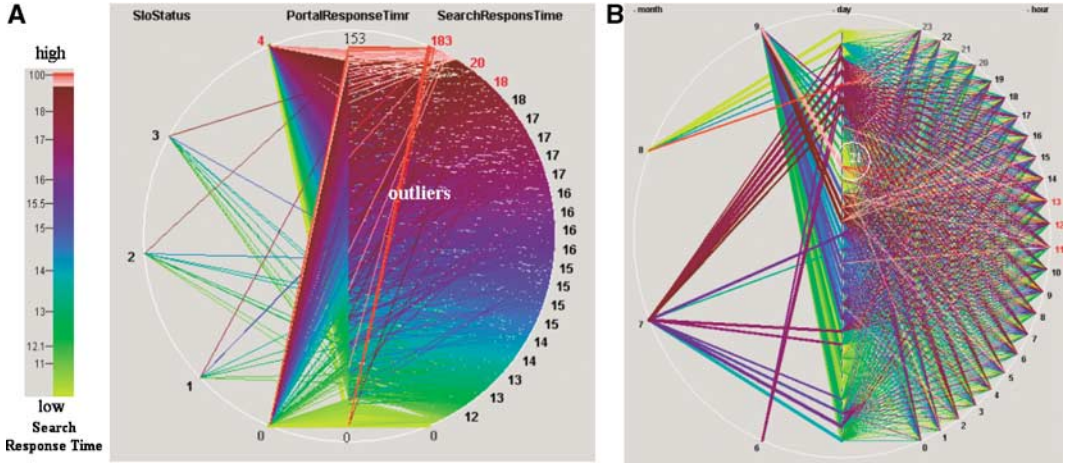

Figure 8 *Service contract process flows* and distribution over time. Portal response time and search response time are highly correlated as seen by nearly parallel lines in (A). Lines with the highest search/portal response times (top 10%) are colored red. Outliers are shown by lines crossing from low portal response time to high search response time). SLO Status 4 has the highest search response time in (A) (more red, pink, burgundy). Month 7 and day 21 have the highest search response time – more blue and burgundy in month 7, most red lines in day 21. High search response times occurred after the 10th day of a month (more red, pink, burgundy, and blue).

seen in the blue and burgundy colors in Figure 9b that show the correlation and (b) in Figure 9c and d the lowest probability SLO status (e.g. SLO status 0) is associated with faster response times, as seen in the yellow and green colors in Figure 9d.

Detection of anomalies among impact factors One of the key functions of *VisImpact* is to detect process flow anomalies including outliers. In Figure 8a, *VisImpact* helps to detect outliers, which are shown as thick red lines drawn from high search response times to low portal response times. After the analyst selects the red line and fades out all unrelated connections, a serious search response time problem (occurring in month 9, day 21) is clearly shown in Figure 10a and b. The analyst can move the pointer to drill down to the detail transaction level to find out that the problem occurred at a time when a search engine was unavailable, which caused a long search time for all previously entered transactions. Using *VisImpact* as a real-time monitoring system, these anomalies can be addressed immediately before the SLO violation probability becomes worse.

Evaluation and comparison

In this section we evaluate the results achieved by *VisImpact* and compare our technique to existing approaches. Of course, it is a difficult task to measure the quality of visualization results, since a definitive and strong set of methodologies for measuring the 'goodness' or the value of a given visualization is still lacking.[18] First attempts for providing quality measurements for scatterplots are provided in Wilkinson *et al*.[19] but there is still a lot of research in this field necessary. Therefore, our evaluation focuses on the comparison of *VisImpact* with other techniques for multivariate data analysis, in particular scatterplots and parallel coordinates.[1]

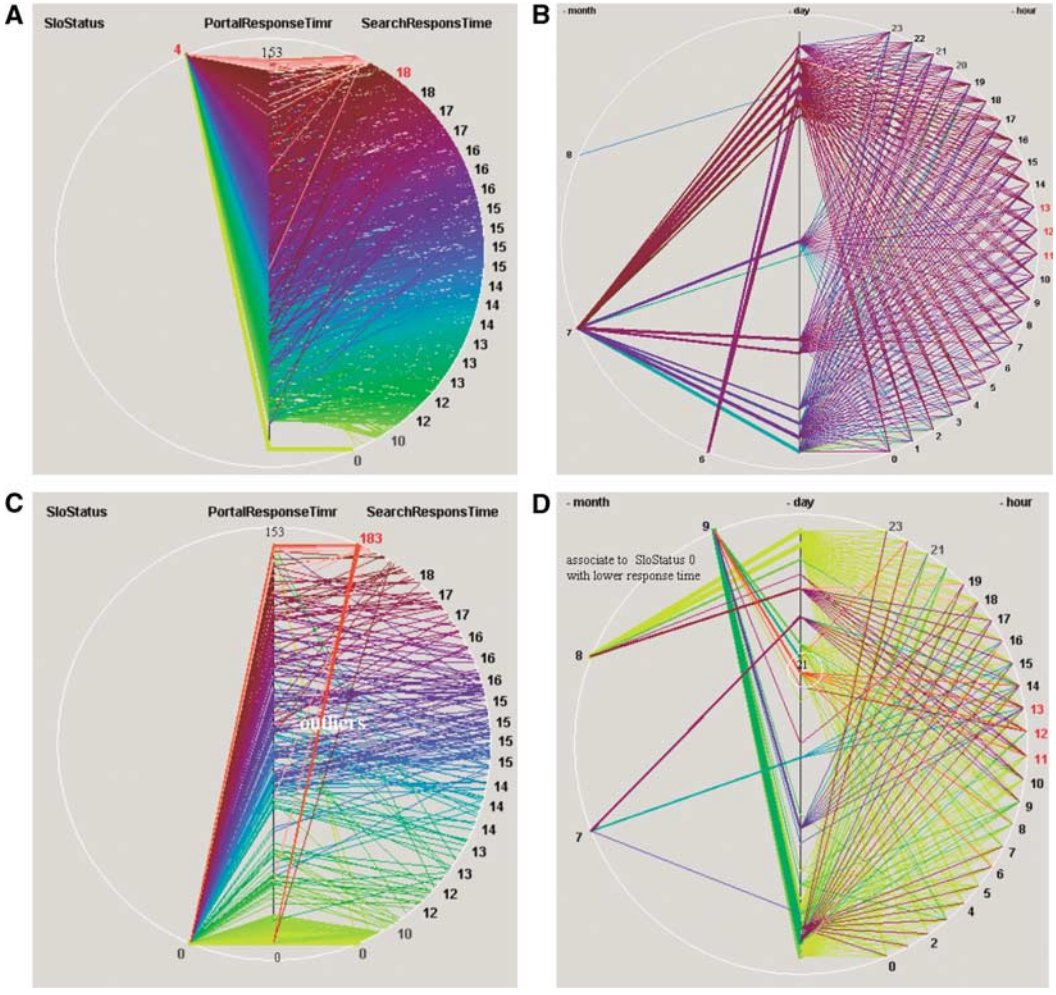

Figure 9 *Process flows and relationships between multiple impact factors*. Graphs are generated when the analyst selects SLO status 4/status 0 in Figure 8a. (A) and (B) show that SLO status 4 is associated with higher response times, as seen by blue and burgundy. (C) and (D) show that SLO status 0 is associated with lower response times as seen by the yellow and green colors. An outlier is detected in (C) and a highest search response time node day 21 in (D).

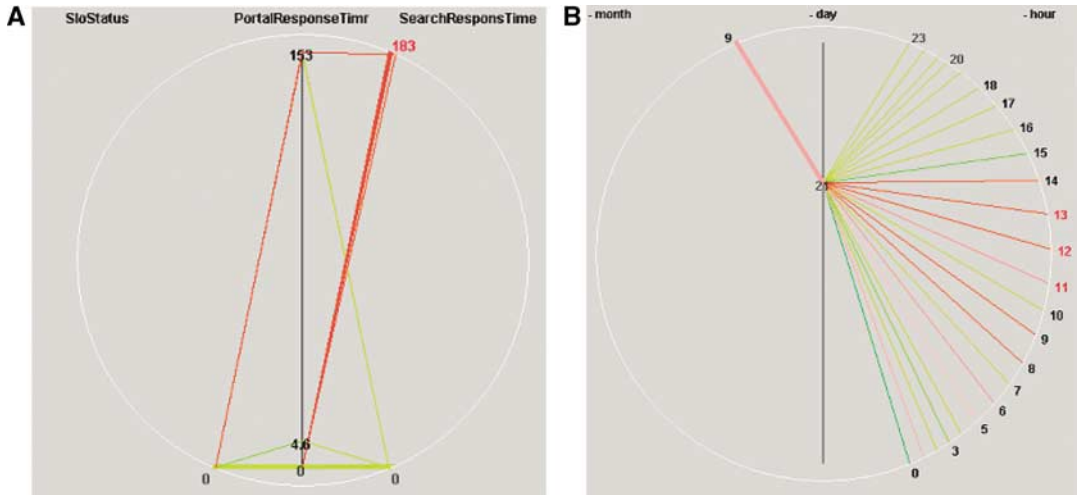

Figure 10 Discover the cause of anomalies (outliers). (A) and (B) are generated when the analyst selects the node on month 9 day 21 (linked with most red lines in (B)). The lines from the anomalies in (A) are linked to the month 9, day 21, and hours (8–14) node in (B). All unrelated lines are faded out for easy identification. The analyst is allowed to move the pointer on the red lines and nodes to display transaction record level information, such as finding server availability in this case.

To investigate the usability of the *VisImpact* approach, future work will include a user study, as first user feedback was very promising. The main advantage of the proposed *VisImpact* approach is the integration of automated analysis techniques into the visualization process to focus on relevant attributes or groups of attributes in the underlying multivariate data sets and thus provide abstract presentations of relevant parts of the data. This allows it to break complex business operations down to groups of relevant business attributes that allows a better business analysis. The analysis results are provided using an intuitive radial visualization layout. The analyst may interact with the system for further data analysis, for example, by using drill-down and roll-up capabilities. To show the benefit of the new technique we used the data set introduced in the section on Service contract analysis and produced scatter-plot matrices and parallel coordinates to analyze the data. Both techniques are commonly used for analyzing multivariate data.

Figure 11 shows a scatterplot matrix of the data set introduced in the section on Service contract analysis. The diagonal contains the attribute names and histograms showing the data distribution. Above the diagonal the scaled absolute correlation values of the corresponding attributes are presented, below the diagonal the data points are shown. Since the scatterplot matrix shows as many scatterplots as there are pairs of parameters, it may be hard for the user to detect relevant patterns if the number of scatterplots is too high.

The figure shows that there is a high correlation between the attributes PortalResponseTime and SloStatus (0.76) and between PortalResponseTime and SearchResponseTime (0.43). Assuming that PortalResponseTime is the measurement of interest and the analyst is interested in attributes which are highly correlated to this measure, he has to construct a new scatterplot matrix that only shows highly correlated attributes, that is, PortalResponseTime, SearchResponseTime and SloStatus. In *VisImpact* this step is done automatically as shown in Figure 8.

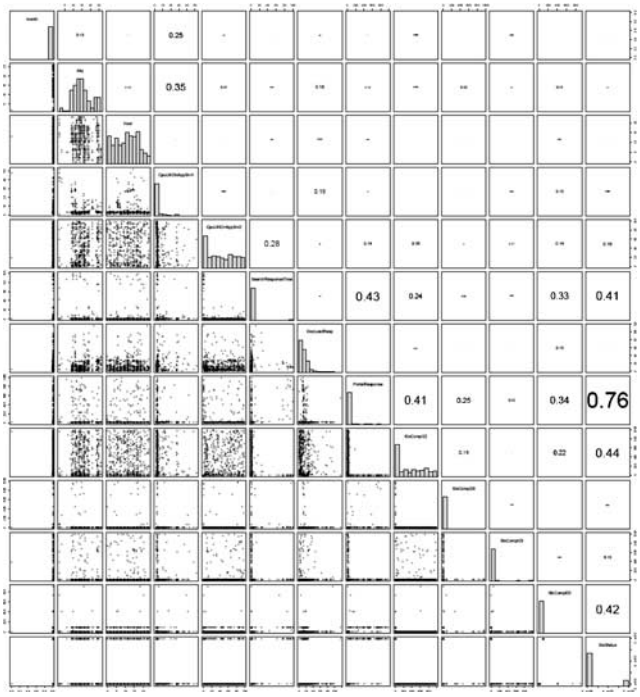

Figure 11 Scatterplot matrix showing a process flow data set and the correlation of it's attributes. In the diagonal, histograms symbolize the data distribution.

Additionally, user feedback indicates that the circular layout in combination with the node ordering options in the *VisImpact* layout provides a much easier way for the analyst to get insight into the relations between the attributes, since owing to overplotting effects scatterplots are often hard to read without additional interaction techniques like brushing.

Another common method for analyzing and visualizing multidimensional data sets are parallel coordinates[1] and their various extensions. Basic idea of this technique is to take all the axis of the multidimensional space and to arrange them in order but parallel to each other. Figure 12 shows the parallel coordinate plot of the business data set introduced in the section on Service contract analysis. Experts in the analysis of such parallel coordinate plots may derive a great deal of data understanding from these plots, but the interpretation can be strongly influenced by the order of the axes.[20] There exist some tools that improve the usability of parallel coordinates plots by providing interaction and analysis techniques like the Xmdv tool,[21] but they do not take the special needs of business analysis processes, described in the section on Basic idea of VisImpact, into account and do therefore not provide the same functionality as the *VisImpact* system. And there is still the drawback that it is hard to identify correlations or clusters between attribute axes that are not located next to each other. Therefore, the visflow graph visualization focuses on three single or aggregated attributes in each single view and provides automated techniques and interaction capabilities for selecting them from the available parameter set.

For example, it is difficult to detect relationships between the three attributes SloStatus, PortalResponse time and SearchResponse time in Figure 12a, therefore the user has to rearrange the axis either automatically based on a correlation measure as shown in Figure 12b or manually. *VisImpact* automates this process, and presents for a given task in each single view only the three most relevant attributes or groups of them and fades out unrelated attributes. Additionally, the circular flow map layout provides more space for placing the data points, for example, according to the weight function, assuming that the diameter of the circle corresponds to the length of a single parallel coordinate axis. Ordering techniques integrated in *VisImpact* additionally help to improve the readability of the resulting visualization for a better visual information representation.

Conclusion

In this paper, we presented the *VisImpact* technique, a new method for business impact visualization. The approach uses a new symmetric circular layout to generate graphs. To simplify the complexity of business operations, *VisImpact* uses correlation analysis, partial matching, cluster, and classifcation analysis techniques to abstract important business impact factors. *VisImpact* presents different impact factors in multiple graphs to view the data from different perspectives. We have addressed the special node placement and coloring problems to make the visualizations useful for discovering patterns and exceptions, and find the cause of business problems by tracing process paths of a transaction record.

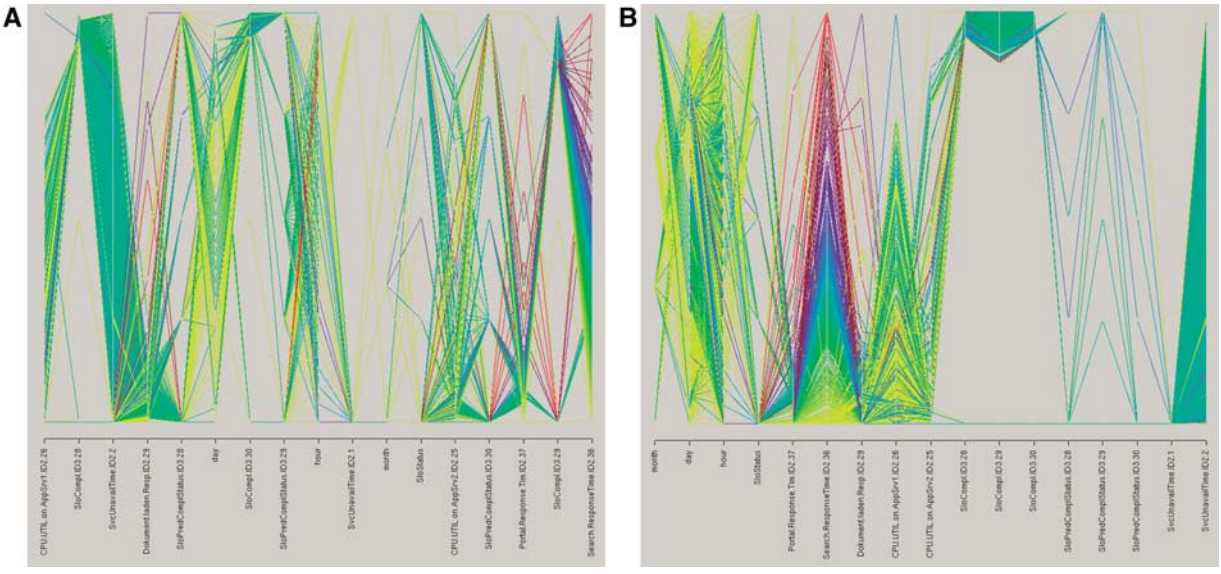

Figure 12 (A) shows a Parallel Coordinate Plot for the Business data example introduced in the section on service contract analysis, Color represents Search Response Time with same colormap as in Figure 8. For non-analysis experts it may be difficult to extract similar information as shown in Figures 8–10 from the Parallel coordinate plots, even with rearranged dimensions according to attribute correlation shown in (B), due to the missing visualization capabilities provided by VisImpact.

We applied the *VisImpact* technique to real data sets from a wide variety of applications, including fraud analysis and service contract analysis. The current experimental studies show significant advantages of the *VisImpact* technique in comparison to existing techniques in performing root-cause analysis. Future work will include the application of the *VisImpact* technique in other areas such as capacity planning and business financial activities.

Acknowledgments

We thank Beth Keer for her encouragement and suggestions; Jörn Schimmelpleng, Stephane Sabiani, Akhil Sahai, Fabio Casati from HP for providing Service Level Management techniques and SLM data; Brian Thesing for the comments and application data, Manish Bhardwaj and Shashidhar Iddomsetty from HP Consulting and Services for their comments and data.

References

1 Inselberg A, Dimsdale B. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: *Proceedings of the 1st Conference on Visualization '90 (VIS '90)*. IEEE Press: Los Alamitos, CA, USA, 1990; 361–378.
2 Eick SG, Steffen JL, Sumner EE. Seesoft – a tool for visualizing line oriented software statistics. *IEEE Trans Softw Eng* 1992; **18**: 957–968.
3 Hao MC, Dayal U, Keim DA, Schneidewind J. Visbiz: a business process visualization case study. In: *Proceedings of the Eurographics/ IEEE-VGTC Symposium on Visualization, EuroVis 2005*. Leeds, UK, 2005.
4 Kaufmann M, Wagner D. *Drawing Graphs: Methods and Models*. Springer Verlag: Berlin, 2001.
5 Chuah MC, Eick SG. Managing software with new visual representations. In: *Proceedings of the 1997 IEEE Symposium on Information Visualization (InfoVis '97)*. IEEE Computer Society: Washington, DC, USA, 1997; 30.
6 Agrawal R, Ling KI, Sawhney HS, Shim K. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In: *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB'95)*. Morgan Kaufmann Publishers Inc: San Francisco, CA, USA, 1995; 490–501.
7 Berchtold S, Keim DA, Kriegel H-P. Using extended feature objects for partial similarity retrieval. *The VLDB Journal* 1997; **6**: 333–348.
8 Mihael Ankerst, Stefan Berchtold, Daniel A.Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In: *Proceedings of the 1998 IEEE Symposium on Information Visualization (INFOVIS'98)* 1998. IEEE Press: Los Alamitos, CA, USA, 52.
9 Yang J, Wang W, Yu P. Mining asynchronous periodic patterns in time series data. In: *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. ACM Press: New York, NY, USA, 2000; 275–279.
10 Faloutsos C, Ranganathan M, Manolopoulos Y. Fast subsequence matching in time-series databases. In: *Proceedings of the 1994 ACM SIGMOD international conference on Management of data (SIGMOD '94)*. ACM Press: New York, NY, USA, 1994; 419–429.
11 Han J, Dong G, Yin Y. Efficient mining of partial periodic patterns in time series database. In: *Proceedings of the 15th International Conference on Data Engineering (ICDE'99)*. IEEE Computer Society: Washington, DC, USA, 1999; 106.
12 Hinneburg A, Keim DA. Clustering techniques for large data sets from the past to the future. In: *KDD '99: Tutorial Notes of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and data Mining*. ACM Press: New York, NY, USA, 1999; 141–181.
13 Hinneburg A, Keim DA. An efficient approach to clustering in large multimedia databases with noise. In: *The Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)* 1998; 58–65.
14 Eades P, Wormald NC. Edge crossings in drawings of bipartite graphs. *Algorithmica* 1994; **11**: 379–403.
15 Sugiyama K. Methods for visual understanding of hierarchical system structures. *IEEE Transactions on Systems, Man, and Cybernetics* 1981; **1**: 109–125.
16 Hao MC, Dayal U, Hsu M, Baker J, Deletto B. A Java-based visual mining infrastructure and applications. In: *Proceedings of the 1999 IEEE Symposium on Information Visualization (INFOVIS '99)*. IEEE Computer Society: Washington, DC, USA, 1999; 124.
17 Sahai A, Machiraju V, Sayal M, Van Moorsel AP, Casati F. Automated SLA monitoring for web services. In: *Proceedings of the 13th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management (DSOM '02)*. Springer-Verlag: London, UK, 2002; 28–41.
18 Miller N, Hetzler B, Nakamura G, Whitney P. The need for metrics in visual information analysis. In: *Workshop on New paradigms in Information Visualization and Manipulation (NPIV '97)*. ACM Press: New York, NY, USA, 1997; 24–28.
19 Wilkinson L, Anand A, Grossman R. Graph-theoretic scagnostics. In: *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'05)*. IEEE Press: Washington, DC, USA, 2005; 21.
20 Robert Spence. *Information Visualization*, 1st edn. ACM Press: Addison-Weslay, 2000.
21 Rundensteiner EA, Ward MO, Yang J, Doshi PR. Xmdvtool: visual interactive data exploration and trend discovery of high-dimensional data sets. In: *Proceedings of the 2002 ACM SIGMOD international conference on Management of data (SIGMOD '02)*. ACM Press: New York, NY, USA, 2002; 631–631.