視覚的データ分析ツールにおけるデータ欠損補完・表示方式

2R－4

岡崎　伸一　磯部　成二
ＮＴＴ情報通信研究所

1　はじめに

データベースからデータの特徴を抽出し戦略的な意思決定に活用するｼｽﾃﾑの構築が望まれている。この要求に答えるため、エンドユーザによる簡易な定義入力だけでデータを自由に視覚化できる視覚的データ分析ツール(ＩＮＦＯＶＩＳＥＲ)[1]を提案した。しかし、本ツールはデータ欠損があるとそれと関連するデータ全体を図形表示できないため、データ分析に支障をきたすいう問題がある。本稿では上記問題を解決するために検討したデータ欠損の検出法及び、補完表示方法について述べる。

2　データ欠損の発生パターン

データの欠損は、人手による投入ミス以外にも、構築途中のデータベースや、複数業務が同時進行して業務毎にデータ投入される場合に発生しうる。データ視覚化ツールにおいては、従属関係のないデータの場合、データ欠損以外のデータは表示されるので問題は少ないが、ネットワーク的なデータの様に関係のあるデータの場合はデータの一部欠損のため

	規　則	制約関係	検出の可能性
単数実体	規則性	有り	○
		無し	×
複数実体	包含関係	収容数既知	○
		収容数未知	×
		収容階層多段	○
		収容階層１段	×
*本稿の対象	連結関係	始終点既知	○
		始終点未知	×
	関係無し	—	×

表１.データ欠損の発生パターン

に表示できないデータが発生するため、データ分析に支障をきたす可能性がある。データ欠損の発生パターンには、表１に示すパターンが考えられる。一般的にデータ欠損はデータ間の規則や制約関係を根拠に検出することがでる。

本稿のデータ欠損検出パターンは表１の複数実体－連結関係有－始終点既知の場合を対象とする。図１に対象とするデータ欠損パターンのデータ構成例を示す。エンドエンド実体としてclient、それを構成する区間実体としてcomponent、各区間の両端実体としてnodeの３実体が本図のように関係している。

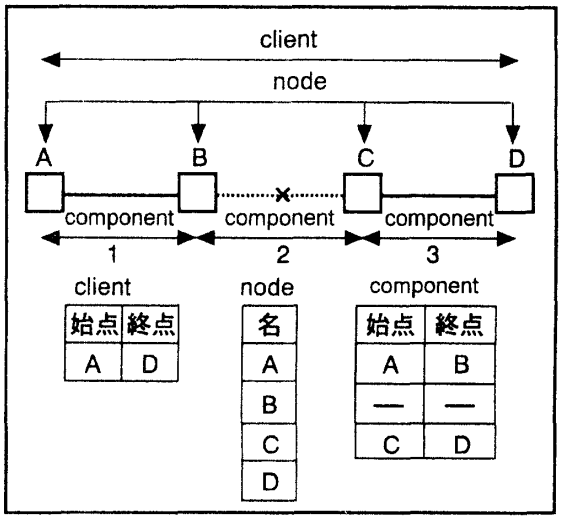

図１.データ欠損パターンの構成例

3　INFOVISERにおけるデータ欠損検出方式

INFOVISERは、文字や数値のデータを集合した実体とその実体間の関係をノード型やライン型の図形オブジェクトとして表示し、視覚的データ分析を支援する。図２に示すようにINFOVISERには、情報源のデータベースから分析対象の実体データを抽出する実体抽出機能と、数値や文字の実体データを図形データに変換する情報変換機能から構成される。

INFOVISERは、実体抽出の過程で、実体間の関係をＯＯＤＢのオブジェクト間リンクとして格納するために、本稿で対象とするような３つの実体

A missing data recovering method
for a visual data analysis tool
Shinich OKAZAKI and Seiji ISOBE
NTT Information & Communication Systems Labs.

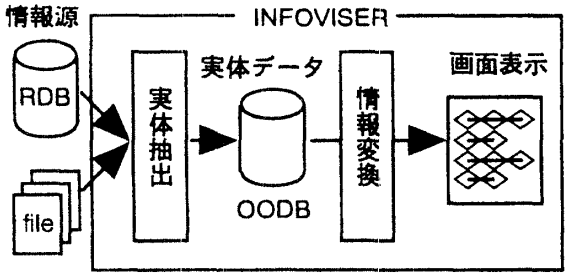

図2.システム構成

(client,component,node) 間の関係を生成している。従って、この生成過程の中に、連結関係の順序性を検査し、データ欠損を検出し、ダミーのデータを補完する機能を追加することにより、以降の図形データへの変換を可能にする方式とした。

4．データ欠損の種類と検出アルゴリズム

4．1　データ欠損の種類

データ欠損の種類は、ノードデータの欠損、ラインデータの欠損に大別される。さらに両者のデータ欠損は、各々、始点、通過点、終点の3箇所の欠損に分類される。データ欠損のパターンは、これら分類の組合せで多くのパターンが存在する。

このような多くのデータ欠損のパターンに汎用的に対処できるアルゴリズムとするために、表2に示すようにデータ欠損の種類を分類し、必要となる補完処理を整理した。

○：欠損データ無し　　始点判定：始点側の欠損判定処理結果フラグ
×：欠損データ有り　　終点判定：終点側の欠損判定処理結果フラグ

		始点	通過	終点	欠損データのイメージ	欠損状態 始点判定	欠損状態 終点判定	補完処理
ライン無し	PT1	×	○	○	◇◇◇◇	1	15	始点ダミーラインの生成
	PT2	○	○	×	◇◇◇◇	6	10	終点ダミーラインの生成
	PT3	○	×	○	◇◇◇◇	6	15	通過点ダミーラインの生成

表2.補完処理（例）

4．2　検出アルゴリズム

（1）連結関係検査の基本処理

データ欠損の有無チェックは、次の処理で実現できる。①clientの始点を出発点に決めて、出発点と合致するcomponentの片端を見つける。②該当componentの両端nodeデータが存在するかチェックする。③componentのもう一方の端点と合致する端点を持つcomponentを見つける。④この操作を、clientの終点に到達するまで繰り返す。

（2）データ欠損検出処理の追加

データ欠損が見つかったらチェックアウトする場合は、基本処理のように片側からの検査で問題無いが、例えば、出発点側のラインが欠損している場合、それ以降のデータの欠損状況が検査できないために、より正確なデータ欠損状態を検出できない。また、データ欠損を1箇所発見しても関連するデータの欠損状態が把握できないと、より最適なダミーデータの生成が難しい。このため、本アルゴリズムは基本処理で欠損の有無をチェック後、データ欠損のあるclient に対し下記の2つの処理を追加した。

1）始点側と終点側から検査を行い欠損状態を記録する。

2）欠損状態を総合的に解析してダミーデータを生成する。

上記のアルゴリズムの概要を図3に示す。

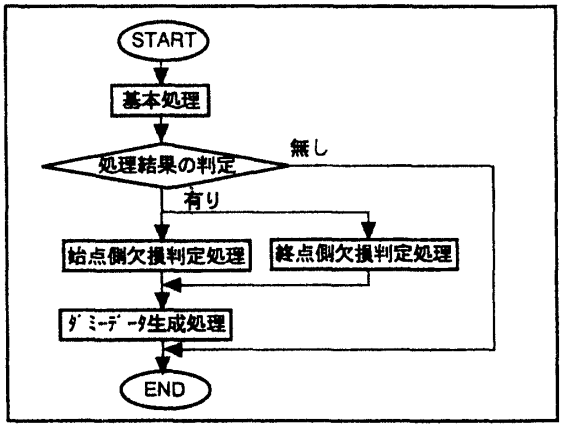

図3.データ欠損検出補完フロー

5　おわりに

本稿で述べたデータ欠損パターン自動判定及び、ダミー図形補完表示手法により、データ欠損がある場合もデータ分析が可能となった。又、データ欠損をユーザに通知し、データ分析への影響を軽減することでデータベース精度向上に貢献することも期待できる。

参考文献

[1]磯部,黒川,塩原 "DB情報ビジュアル化技術" NTT R&D,vol45,No1,1996

[2]塩原,磯部,黒川 "エンドユーザによるネットワーク情報のビジュアル化環境" 電子情報通信学会IN研究会,IN95-150,1996.