

あらまし POS データベースに代表される巨大数値データベースから AI の基礎データとなる学習用教師データを作成することが求められている。巨大な数値データベースはその概略の理解や構造把握が困難であり、学習用データとする以前の構造理解が大きな問題となっている。本研究では、AI 学習用データとして数値データベースを活用するためのツールとしてデータブラウジングを目的とした分析ツールのアーキテクチャを提案する。提案するシステムは演算要素としてデータベース走査と可視化環境を持つスタックアーキテクチャの仮想マシンを定義することにより、多様な可視化処理プログラムを短時間で作成することを可能にしている。

キーワード インターネット、分散情報記録、数値データベース、可視化、ブラウザ、データマイニング、AI

1. まえがき

ビジネスインフラのIT化が進んだことに加え、大容量ファイルシステムが一般化したことにより、情報の自動蓄積による巨大データベース構築が各分野で進んでいる。たとえば、スーパーマーケットやコンビニショップの売上げ情報、いわゆるPOS(Point of Sale)情報はそういった巨大データベースの典型であり、在庫管理だけでなく、統計分析やデータマイニングによって消費者の購買動向などの分析、予測に使われている[1]。POSデータのような巨大データベースは一般公開されることがまれであり、その取り扱いに関する技術的検討、利用モデルの研究はその情報にアクセスできる組織内で閉鎖的に行われていることが多かった。

近年、一部でPOS情報を有料で公開する情報サービスが始まったなど、今後このような巨大データベースが利用可能になりそうな動きがあり、その多角的な応用技術に関する研究活動が期待されている。その応用範囲は経営分析にとどまらず、その可視化結果は一種のデジタルコンテンツとして興味深いものである。しかし、POS情報のような巨大データベースを自由に分析して画像化、映像化する手法についてはほとんど公開された研究が見当たらず、未開拓の分野であるといえる。本稿では、年間売上数千億円規模の大規模店舗のPOSデータを具体例として取り上げ、ブラウジング、可視化という視点で取り扱うための処理アーキテクチャを提案するとともに、試作した可視化環境とそれによるブラウジング感覚のPOS情報可視化について紹介する。

2. POS 情報の構造と規模

POS 情報は売上げ集計端末 (POS レジスタ) に入力された売上げ情報をデータベース化したものであり、最も粒度の細かいレベルでは顧客が受け取るレシートに記載される全情報が記録される。年間売上が 1000 億円のスーパーマーケットチェーン店全体のデータベース規模を概算すると次のように推定できる。スーパーマーケットでの商品平均単価を 200 円と仮定すれば、年間売上げ商品数は約 5 億点と見積もれる。売上げレコードの構造は、売上げ日時、店舗コード、商品コード、分類コード、メーカーコード、顧客コード、単価、個数

共同研究報告書

研究課題：可視化を応用する AI 学習用データ作成の高速化の研究

研究担当者 北海道大学大学院情報科学研究科

特任教授 山本 強

など 8 フィールド以上で構成されており、仮に各フィールドを 4byte と定義すると 1 レコードあたり 32byte 程度の固定長レコードで表現されることがわかる。これが 5 億レコードであるから、全体では $5 \text{ 億} \times 32 = 160 \text{ 億 byte}$ 、つまり POS 情報の内部表現は 16Gbyte 程度の数値データベースであることが予想できる。生の POS データは流通業の経営データそのものであるから、それがそのまま公開されることはなく、顧客情報や時間精度を落とした 2 次データとして公開されることになる。私が共同研究素材として使用許諾を受けているマスタデータは商品単位で、商品分類コード、店舗コード、売上日、売上個数、売上額の情報を 1 レコードとしたものである。図 1 にレコード構成を示す。ちなみに、年間 1560 億円の売上げに対して、レコード数は約一億強である。

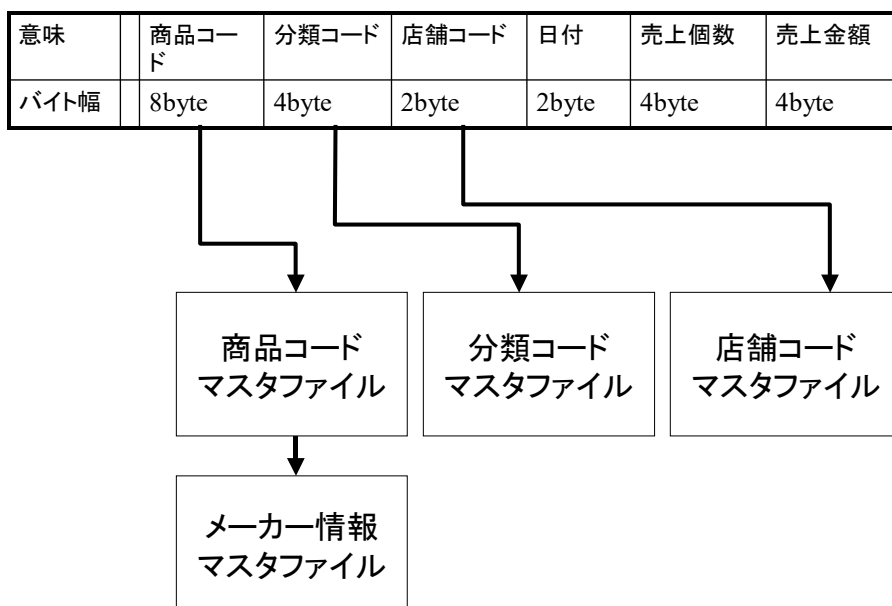


図1 想定するPOSデータベースの基本構造

この POS データベースは関数モデルで抽象化して考えることができる。変数が商品コード(item)、店舗コード(shop)、日付(date)の 3 変数、値が売上数(count)と売上額(sales)の 2 値をとる多値関数とみなすことができる。

$$\text{Sales} = f_1(\text{item}, \text{shop}, \text{date}) \quad (1)$$

$$\text{Count} = f_2(\text{item}, \text{shop}, \text{date}) \quad (2)$$

POS データは数値データベースであり、仮にそれを入手しても意味が全くわからない。その意味を理解するためには、店舗コード、商品コード、分類コード、メーカーコードなどのマスターファイルが必要であり、それと数値データベースを関連付けて初めて意味がわかることとなる。

3. POS 情報可視化の基本方針

POS データの可視化、グラフ化は商用システムで既に行われているが、それはその分野での標準的な集計手法や定番のグラフを作成するものがほとんどで、一般化された可視化手法ではない。根底にあるのは、データベースデータベース本提案では、数値データ可視化の標準アーキテクチャを提案する。

前章で説明したように POS データは多変数多値関数と抽象化することだでき、可視化処理は多変数関数を最終的に 2 次元以下の関数に変換し、グラフや幾何情報に変換することである。POS 情報の軸(フィールド)をなんらかの計算処理で低次元化し、その後レンダリング処理により画像化することで行われる。ここでは低次元化処理を縮退処理と呼ぶこととする。

縮退処理は多変数関数の積分処理と同じことである。例えば、一定期間についてある商品(コード N)の売上額のグラフを作るという縮退処理は以下の積分計算を行うこととなる。

$$G(N, \text{date}) = \int f_l(\text{item}, \text{shop}, \text{date}) d\text{shop} \quad (3)$$

同様に、ある店舗(S)の特定分類(C)に属する商品の年間売上げランキングを集計するという処理は

$$\text{sort}[L(I \in C, S)] = \int f_l(\text{item}, \text{shop}, \text{date}) d\text{date} \quad (4)$$

という計算処理になる。つまり、データベースを多値他変数関数と抽象化し、積分(縮退)とソートを基本演算とすることで基本的な集計処理の演算集合を網羅できるようになる。

提案するアーキテクチャではPOSデータを任意の軸を縮退することで最終的に2次元あるいは1次元のグラフオブジェクトに変換して表示する。どの軸(フィールド)を縮退するかで、多種のグラフ画面を作成することができる。図2は提案システムの可視化画面の例であるが

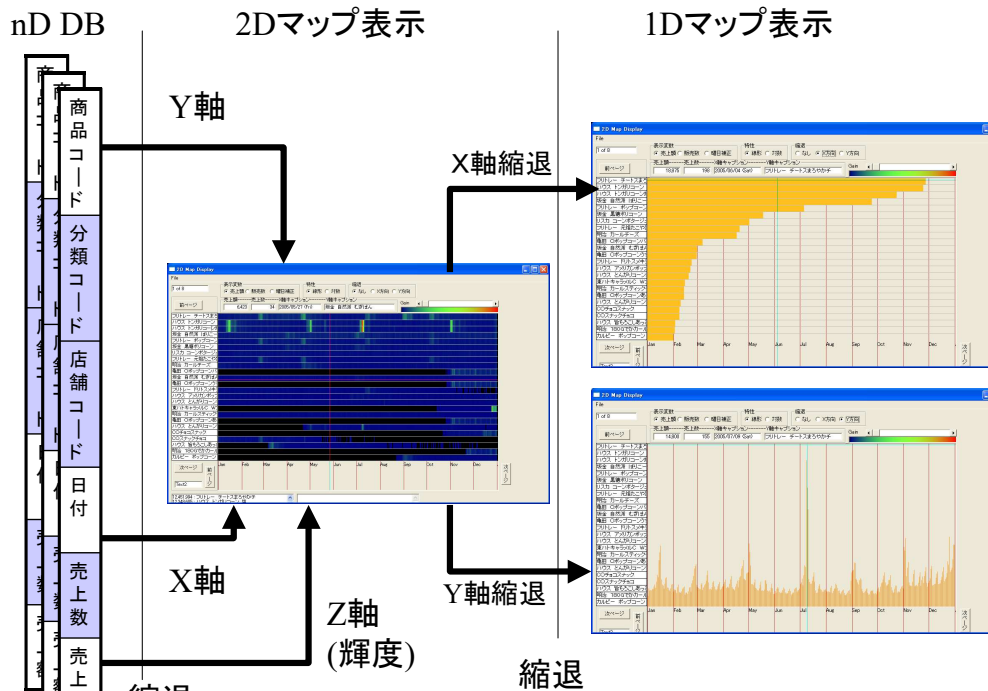


図2 縮退演算による可視化処理の流れ

4. スタック指向の仮想マシンアーキテクチャ

POS データ可視化はデータベースの縮退と画像化処理の組み合わせで行われる。従来の代表的可視化アーキテクチャ[2]は複数の処理をデータフローモデルで結合するものが多い。データフローモデルは3次元画像生成のグラフィックスパイプラインモデルのアナロジーがあり、ハードウェア集合体として可視化システムを抽象化している。これはコンピュータが発明された当時にENIACが演算ユニットのパッチボード接続により計算を定義していたのと同じモデルである。その後、コンピュータアーキテクチャが、いわゆるノイマン型のスタックマシンに基本構造を変化させてきたことを考えると、可視化システムの基本構造をよりフレキシブルな仮想マシンアーキテクチャで構築するほうが良い可能性もある。

ここで提案するアーキテクチャは縮退演算の結果を2値1変数の正規化されたチャートオブジェクトとし、それを一時的に格納するスタックを介在させることにより、コンパクトな可視化アーキテクチャを実現する。図3に提案するアーキテクチャの基本構造を示す。

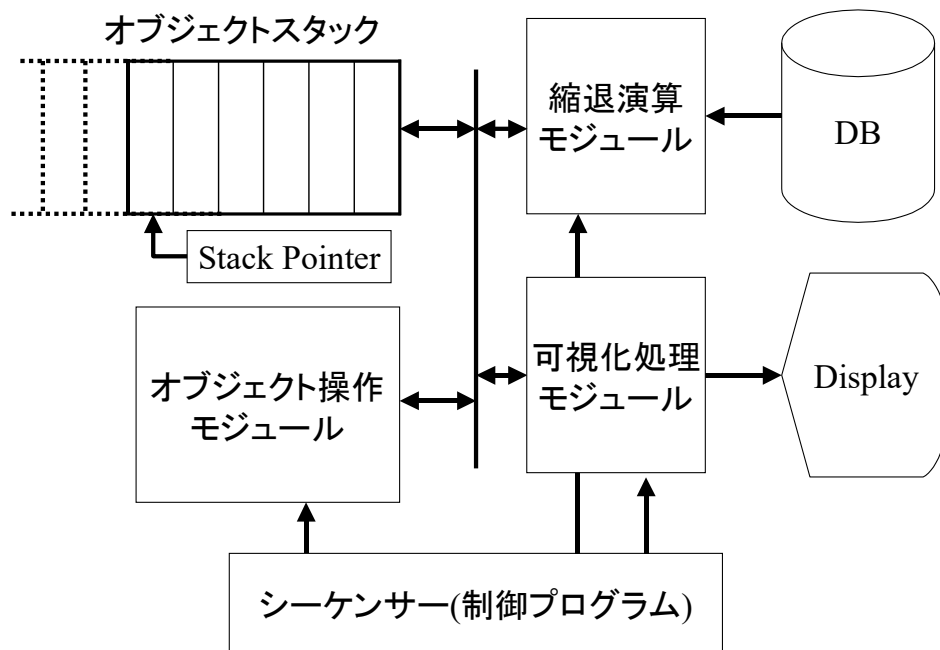


図3 スタックを併用する可視化処理用仮想マシンアーキテクチャ

4.1 スタックを経由するオブジェクト操作

スタックに対する基本操作として以下の代表的な操作命令が考えられる。

移動操作	PUSH (表示中オブジェクトをスタックに移動)
	POP (スタックを1段POP)
演算操作	加減乗除 (スタックの先頭の2オブジェクトに対する2項演算)
	縮退 (スタック全体を加算した結果のみをスタックに残す)
	正規化 (スタックの先頭オブジェクトをスタックの一番底にあるオブジェクトで除算する)
展開操作	展開 (変数を2まで縮退した後、1変数についてスタックに展開する)
初期化操作	クリア (スタックを初期化する)
定数生成操作	定数 (値が定数、固定関数であるオブジェクトを生成し、スタックにPUSH)

シーケンサはデータベース縮退とスタック操作命令を組み合わせ、順序実行することで高度な可視化処理を実現することができる。

4.2 マニュアル操作による可視化プロセス合成

可視化処理にスタックを取り入れることの利点はソフトウェアの構造的な整理だけでな

共同研究報告書

研究課題: 可視化を応用する AI 学習用データ作成の高速化の研究

研究担当者 北海道大学大学院情報科学研究科

特任教授 山本 強

く、マニュアル操作による可視化操作にもある。これは、かつて Hewlett Packard 社, Texas Instrument 社のプログラム電卓の基本操作（逆ポーランド記法[3]）と似ている。シーケンスの制御プログラムをマニュアル操作で置き換えることにより、高レベルな電卓操作と似た感覚で可視化処理を行うことが可能になる。マニュアルでの操作をジャーナル記録することで、その操作を新しい可視化プロセスとすることもできることになる。

従来、POS 情報の分析はデータベースに対してあらかじめ決められた集計処理手順を行い、結果を表示するという形式のものが多かったが、ここで提案するモデルでは、集計手順をモジュール化しその中間処理結果をスタックに複数保存してそれらを演算処理して様々な表示形態を作成することができ、分析者主体で表現を作成することが可能になる。

5. 実装例

提案したバーチャルマシンを元に、POS データ可視化およびブラウジングを目的としたソフトウェアツールを開発したのでその概要を紹介する。開発したシステムは 1 億以上の POS データレコードを短時間で縮退・画像化することを目的としており、データベースエンジンで POS データの構造に最適化することにより、極めて短時間で縮退処理を実行可能にしている。高速化はデータベースの前処理により達成されており、POS データのリアルタイム処理ではないが、一日単位の集計に対する差分処理は短時間で完了するので、実用的には問題はない。

本システムのもう一つの特徴は、可視化された商品、分類、店舗などの関連情報がハイパーテキスト的にリンクされていることで、POS データをブラウズする感覚で分析することができる。図 4 に実装したシステムでの POS 情報ブラウジングのイメージを示す。可視化結果の表示モジュールは何種類か作成されているが、それらに表示される各種情報は可能な限りそれが自然に表現される表示モジュールに遷移できる構造を実現している。

共同研究報告書
研究課題：可視化を応用する AI 学習用データ作成の高速化の研究
研究担当者 北海道大学大学院情報科学研究科
特任教授 山本 強

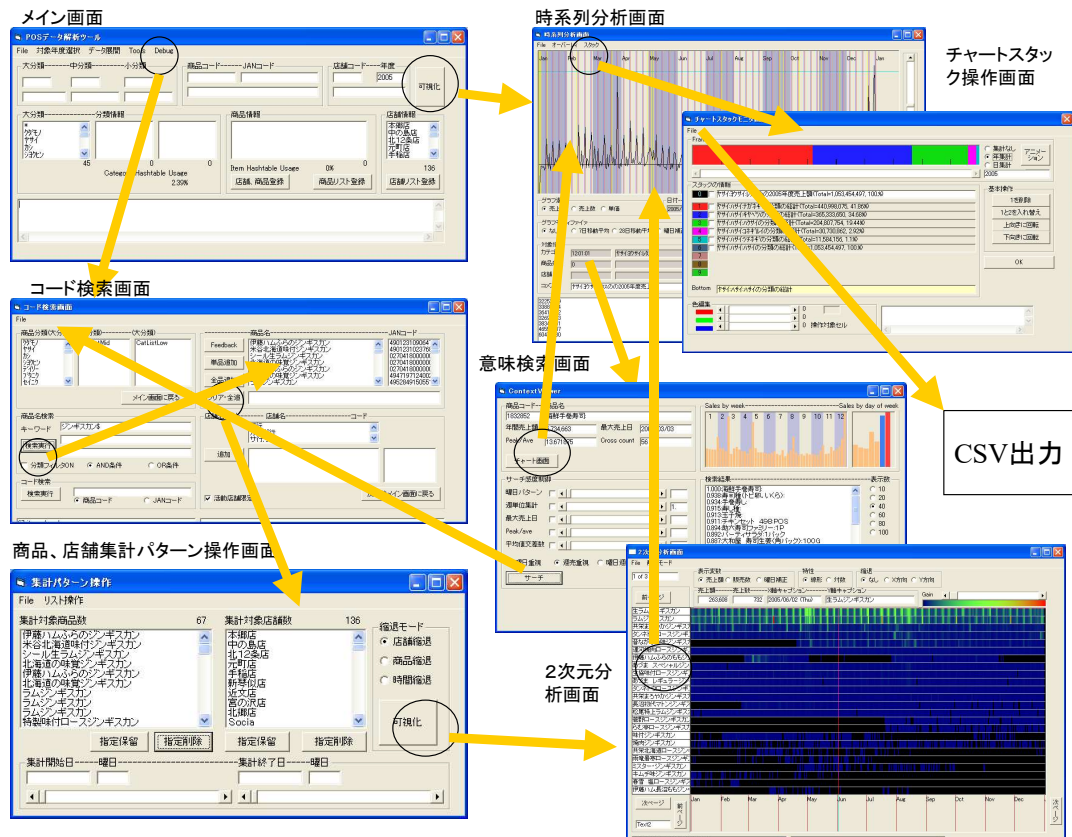


図4 実装したPOS情報可視化システムの代表的表示モデルと遷移関係図

6. むすび

スタックマシンモデルで仮想マシンを定義し、それを使った数値データベース可視化システムを提案し、実装した。提案したシステムではブラウザで閲覧する感覚で巨大なPOSデータベースを可視化することができ、POSデータに現れる様々な消費傾向や地域性を発見することができた。ITインフラの強化により、自動収録される巨大データベースは今後増えてくると予想されるため、データブラウザの汎用化、高機能化を進めることが重要である。