システムログの俯瞰と縮約のための可視化の一手法

林 亜紀† 伊藤 貴之†

†お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻

1.概要

時系列データを可視化する研究は数多く行われているが，我々はその中でも，大容量で多次元の時系列データに着目している．クレジットカードの決済情報や Web ページのアクセスログといったシステムログなどである．決済情報の場合，我々が入手したデータでは，半年分で 16 万個の決済が記録され，属性はカード番号，加盟店コード，不正種別，商品コードなど 40 個に及ぶ．また，アクセスログについても，著者らが所属する研究室のアクセスログでは，半年分で 20 万アクセスが記録され，属性は IP アドレス，URL，ステータスコードなど 12 個ある．システムログの時間的特徴を観察することで，決済情報では決済の傾向，不正決済特有の傾向を観察できる．また，アクセスログでは，アクセス傾向や，エラーの観察，更新時期の決定などが期待される．しかしながら，膨大な情報から短時間で有意性のある知見を得るのは難しい．

本報告では，各属性に対する時間的特徴変化の俯瞰・縮約による全体像把握，効率的な属性選択の支援を行う Visual Analytics Tool を提案する．俯瞰表示では X 軸を時系列，Y 軸を各属性値(項目)として，各時刻・項目の集計値をヒートマップ表示する．加えて，ソートによる縮約と，クラスタリングによる縮約を行う．また，各属性に対する推薦度を算出・提示することで，有意性のある結果を促す属性推薦も実現する．

2.関連研究

一般的な多次元データ可視化のための属性選択推薦は，散布図など特定の可視化手法を対象として実現されている[1][2]．時系列に着目した手法[3]では，効果的な結果の推薦は考慮されておらず，一次元時系列データで実現されているような縮約機能との融合例も少ない．本手法では，多次元データの時系列に着目し，結果の有意性を尺度とした属性推薦や，2 種類の縮約を実現する．また，システムログの分析システム WireVis[4]は複数の可視化手法の融合によるが，[2]の散布図で時系列以外の 2 属性を可視化したものと本手法とで，[4]に近い効果が得られる．属性選択の幅が本手法よりも広い[2]の散布図と，本手法との併用から，新たな発見が期待される．

3.提案内容

3.1 時間的特徴の俯瞰表示

X 軸を時系列，Y 軸を属性値(項目)，色を該当件数として，各座標に集計結果を描画する．X 軸は日付，曜日，月，時間の 4 種類から選択でき，Y 軸には時系列を含む全属性が選択できる．暖色ほど集計値が大きいことを示す．クリックで項目名，集計値を表示する．属性値で描画するログを限定するフィルター機能も実現する．

3.2 Y 軸選択の推薦機能

有意性のある可視化結果を得られる属性を推薦する機能を提案する．選択された以下のいずれかの値を相対的に算出し，各属性の推薦度をインタフェース上のボタン色の濃度で提示する．

(1) 全体の最大値→特異な集計が存在
(2) 各項目の最大値の平均→特異な集計が頻発
(3) 項目の合計値の最大値→特異な項目が存在
(4) エントロピーの低さ(ちらばりの大きさ)
(5) エントロピーが高さ(ちらばりの小ささ)

3.3 可視化結果の縮約

3.1 の俯瞰表示は，例えば URL が数千～数万種類におよび，属性によっては画面上の水平な帯が増え，拡大操作でちらつきが見られる場合がある．そこで，2 種類の縮約機能を提案する．

3.3.1. ソートによる縮約

以下のいずれかの値を項目ごとに算出してソートし，スライダーで表示項目数を調節する．

(1) 各項目の最大値
(2) 増加量の最大値
(3) エントロピーの低さ(ちらばりの大きさ)
(4) エントロピーの高さ(ちらばりの小ささ)

3.3.2. クラスタリングによる縮約

ソートによる縮約では，一定の周期性がある項目など，有意性があるにも関わらず描画されない項目も出てくる．そこで，項目をクラスタ

"A Visualization Technique for System Logs Realizing Overview and Compression",
Aki Hayashi † , Takayuki Itoh †
†Graduate School of Humanities and Sciences, Ochanomizu University

リングする縮約を提案する．k-means 法で 30 個程度にクラスタリングした後，各クラスタの代表を描画することで縮約する．加えて特定のクラスタに属する項目全体の描画も可能である．

4.実行結果

4.1 決済情報の傾向分析

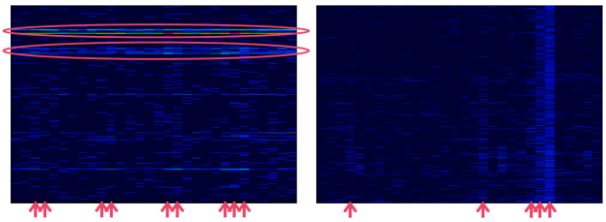

図 1：2007/11 の決済情報の傾向分析例

図 1(左)は X:日付，Y:加盟店とした俯瞰表示結果である．赤丸の特定加盟店で継続的に決済が多いことや，休日にあたる矢印の日に多店舗で周期的な決済増加があることが分かる．図 1(右)はクラスタリングを適用後，特定のクラスタを描画した結果である．月末の連休最終日に決済が多い加盟店の多くで，連休の他の日や，他の休日にも決済が多いといった特徴を観察できる．

4.2 不正の傾向分析(2007/7-12 月の不正決済)

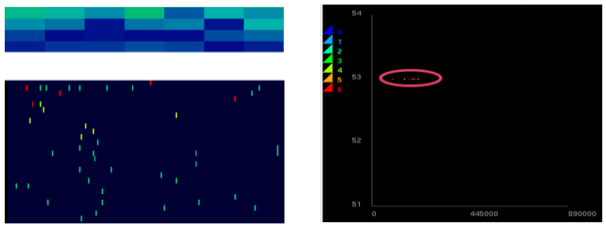

図 2：2007/7-12 月の不正決済の傾向分析例

図 2(左上)は X:曜日(左が日曜)，Y:商品コードとした結果の一部である．海外商品(1 行目)は継続的に多いが，電化製品(2 行目)や鉄道(3 行目)は土日に多く，通販(4 行目)は平日に多い．図 2(左下)は電化製品に限定して X:日付，Y:推薦(3)より加盟店コードとして，ソート(1)で縮約した結果である．2 行目の加盟店で継続的に不正が見られる．Y:カード ID とすると，同一 ID ではなかった．さらに図 2(右)で 2 行目の加盟店を[2]の散布図で観察した．X:金額，Y:不正種別，色:曜日としたところ，土日に 10 万円前後の偽造による不正が集中しており，この加盟店が偽造カードの標的となっている可能性が浮かび上がった．

4.3 アクセス・エラーの傾向，更新時期の検討

図 3(左上)は，X:月，Y:URL として，(1)でソート・縮約を行った例である．元の項目数は 6000 個であり，縮約により可読性が大幅に向上した．アクセス数が多いことが自明な top や index に加えて，教員のある授業全 11 回のうち，7,8 回目の資料だけアクセス数が多いことが分かった．図 3(左中央)で，7 回目の資料について，X:時間(左が 0 時)，Y:曜日(下から日，月…)としてアクセス時期を観察した．月曜午前の授業に向けて，日曜夕方から月曜朝にアクセスが多く，特に日曜の 23 時にアクセスが集中しており，更新はそれ以前に行う必要があることが分かった．続いて，図 3(左下)で，X:月，Y 軸:推薦(4)よりステータスコードとしたところ，最も一般的な成功(200)よりも，部分的成功(206)の方が多いことが分かった．この結果からアクセス増の原因がこの 206 ではないかと考えられる．続いて図 3(右)で，全授業資料のアクセス数を X:URL，Y:ステータスコード，色:転送量，高さ:件数として[2]の散布図で観察した．予想通り 7,8 回目の授業のステータスコード 206 のアクセス数が突出している．観察を続けると 206 が発生した IP アドレスは多様であり，多くの学生が 7,8 回目の資料の入手に苦労している可能性が発見できた．

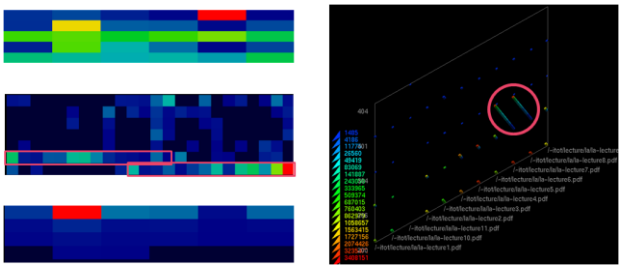

図 3：2011/4-9 月のアクセスログ分析例

5.まとめと今後の課題

本報告では，大容量で多次元の時系列データであるシステムログを効率的に俯瞰・縮約表示する可視化手法を提案した．今後は各機能の改良に加えて，項目をまたいで類似した特徴変化パターンを提示する機能の実現を検討している．

謝辞：決済情報を提供いただきました株式会社インテリジェントウェイブ様に感謝いたします．

参考文献

[1] A. Tatu et al., Combining Automated Analysis and Visualization Techniques for Effective Exploration of High-dimensional Data, IEEE Symposium on Visual Analytics Science and Technology, pp. 59-66, 2009

[2] C. Sakoda et al., Visualization for Assisting Rule Definition Tasks of Credit Card Fraud Detection Systems, IIEEJ Image Electronics and Visual Computing Workshop, 2010.

[3] W. Aigner et al., Visualizing time-oriented data - A systematic view, Computers and Graphics, Vol. 31, No. 3, pp. 401-409, 2007.

[4] R. Chang et al., WireVis: Visualization of Categorical, Time-Varying Data from Financial Transactions, IEEE Symposium on Visual Analytic Science and Technology, pp. 155-162, 2007