時系列データの俯瞰と縮約のための可視化

Time-series Visualization adopting Overview and Summarization

林 亜紀♥

Aki HAYASHI

本論文では様々な時系列データを対象として，全体像を把握する俯瞰と，重要部分を詳細観察する縮約を実現する可視化手法を提案する．近年，多様な時系列データが蓄積されている．例えば Web のアクセスログや決済情報などのシステムログ，SNS を通した位置情報ログなどのライフログ，MIDI に代表される楽譜情報など時系列データは多種多様である．時系列データは特に多属性かつ大規模なため，俯瞰と縮約を実現した可視化手法は少ない．本論文では，一般的なシステムログ，独自の表示形式を持ったライフログや楽譜情報それぞれについて，データが持つ周期性や統計量に着目した自動抽出指標を導入することにより，俯瞰と縮約を実現する効果的な可視化手法を提案する．

This thesis presents visualization techniques for various kinds of time-series data adopting overview and summarization. Recently, many kinds of time-series data are accumulated, for example system logs such as web access logs and transaction logs, life logs including location-based information obtained from social networking services, music data including score information such as MIDI. Such time series data are so massive that effective visualization utilizing interaction techniques realizing overview and summarization thought to have been difficult. This thesis proposes sophisticated visualization techniques adopting overview and summarization by employing automatically calculated indices considering periodicities and statistics. The thesis introduces three visualization techniques for general system logs, life logs and music score data: latter two have distinct visualization styles.

1. はじめに

近年，様々な時系列データが蓄積されている．Webのアクセスログ・決済情報などのシステムログや，ソーシャルネットワーキングサービス上で蓄積される位置情報ログなどのライフログに加え，音楽メディアをテキスト化したMIDIも時系列データの一種である．それらが持つ時間的周期性に着目した分析により，将来予測や状況・内容理解の精度が向上し，リコメンデーションやマーケティングなどの経済効果が期待される．

形状や空間構造をもたない抽象的なデータに対し，形状を与えて提示することにより，人間による視覚的な理解を支援する技術として，情報可視化への注目が高まっている．特に，既存の可視化手法を効率的に活用して自然現象や社会現象を効率的に分析するVisual Analyticsの有効性が示されてきている．一般に，Visual Analyticsでは，まずデータ全体の傾向を俯瞰した上で，情報を縮約して興味深い情報を詳細観察する，といった俯瞰・縮約の分析手順が提唱されている．しかしながら時系列データは大容量かつ多属性なため，効果的な俯瞰・縮約を実現する可視化手法は少なく，先行研究である文献[1-3]においても，データの値そのものを拡大・縮小表示したり，簡単な統計を表示したりといった俯瞰・縮約にとどまっており，縮約箇所の選択は分析者の技量に依存する．

本論文の目的は，大規模時系列データに対し，データ特性や観察目的に応じた効果的な俯瞰・縮約を実現することである．特に，観察支援のための自動抽出指標の導入により，可視化の可読性を向上させるとともに，俯瞰表示の中から縮約による詳細観察にふさわしい興味深い部分を抽出する作業を支援することに注力する．本論文では，一般的な多次元時系列データを扱う一般的な手法に加え，前述した位置情報データや楽譜データのような，独特なデータに対しても対応した網羅的な俯瞰・縮約のための可視化手法の構築を目指して，3つの手法を提案する．

2. システムログの俯瞰・縮約

2.1 提案手法概要

本章では，決済情報やアクセスログなどのシステムログに代表される，一般的な多次元時系列ログの俯瞰・縮約のための手法を提案する．図 1 に提案するインタフェースの画面例を示す．システムログの分析目的として，サイト運営やマーケティングなどの経営支援が挙げられる．

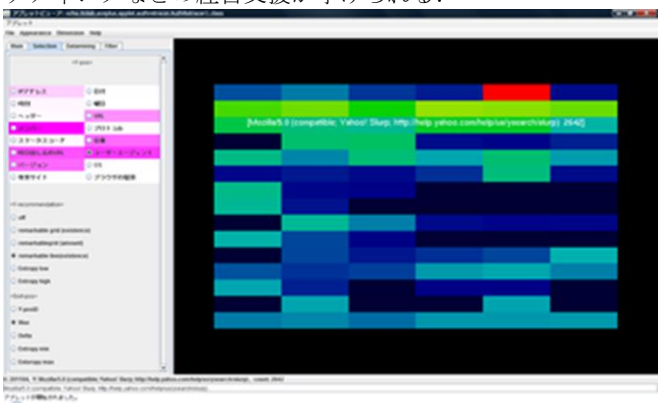

図 1　システムログの俯瞰・縮約インタフェース

・ヒートマップ切り替えによる俯瞰

大規模かつ多属性で，俯瞰が難しいシステムログを，高い可読性を保ちながら可視化するために，図 1 の右画面のように X 軸を時系列，Y 軸を選択された 1 属性の属性値(項目)，色を該当件数として，各座標に集計結果を描画したヒートマップを切り替えながら観察する可視化方法を採用する．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．

・属性選択推薦

♥ 正会員　日本電信電話株式会社 NTT サービスエボリューション研究所 aki@itolab.is.ocha.ac.jp

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実現する．図 1 の画面左の操作パネルで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（桃色）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)

・可視化結果の縮約

前述の俯瞰表示は，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，ソートによる縮約とクラスタリングによる縮約の 2 種類の縮約機能を提案する．

ソートによる縮約では，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) 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)

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が出現してしまう場合がある．そこで，項目をクラスタ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．

2.2 可視化例

決済情報とアクセスログの実データを用いた分析により手法の有効性を確認し，一般ユーザとデータ分析に長けた専門家の両方から実用面での評価を得た．ここでは，各データについて可視化例を 1 つずつのみ紹介するが，その他の可視化例については博士論文及び文献[4]を参照されたい．

・決済情報を用いた不正観察例

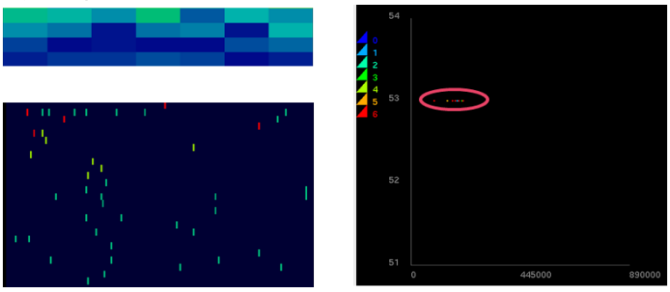

図 2　2007/7-12 月の不正決済の傾向分析例

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．

・アクセスログを用いた傾向観察例

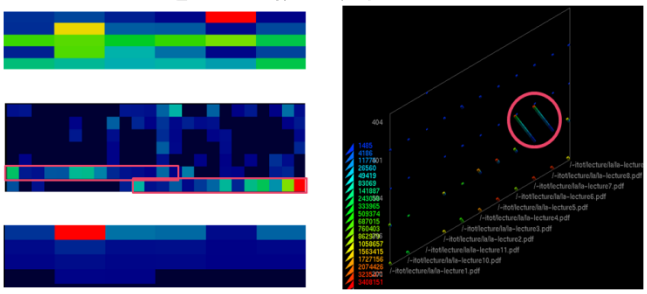

図 3　2011/4-9 月のアクセスログ分析例

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．

3. 位置情報ログの俯瞰・縮約

3.1 提案手法概要

本章では位置情報ログに代表されるライフログの俯瞰・縮約手法を提案する．ライフログの分析目的として，情報推薦の満足度向上が挙げられる．しかしながら，ユーザニーズに合った推薦内容とタイミングの決定は難しい．ユーザの状況は時々刻々と変化し，その状況を詳細に抽出するのが難しいためである．本論文では，ライフログの様々な時間軸に基づく周期性に着目し，各行動がどの程度本人の習慣に沿っているのかを自動で定量的に評価する習慣度という指標を導入する．例えば旅行中などの非習慣時には，他の旅行者が現地で訪れやすい観光スポットを推薦するなどの活用が考えられる．本章では位置情報ログとしてチェックインを想定する．

・習慣度の定義

提案する習慣度では，例えば曜日や時間帯，その両方や時間考慮なしなどの複数の時間軸を考慮して，各行動がどの程度習慣に沿っているかを数値化する．まず，軸毎に訪問確率分布$M_{lut}^{(i)}$ を定義する．ここで i は時間軸，l は訪問場所，u はユーザ，tは訪問曜日や時間帯を表す．次に，以下の式で

時間軸毎の習慣度 $\mathcal{R}_{lut}^{(i)}$ を算出する．

$$\mathcal{R}_{lut}^{(i)} = \frac{M_{lut}^{(i)} - \mu}{\sigma}$$

ここで，σ は標準偏差，μ は平均を示し，$\mathcal{R}_{lut}^{(i)}$ は訪問確率分布中の突出度に該当する．続いて，以下の式により，時間軸ごとの習慣度の重み付線形和として，総合的な習慣度を定義する．重み$w_{ut}^{(i)}$はtにおけるユーザuの相対的なログ数を考慮して自動的に決定する．

$$\mathcal{R}_{lut} = \sum_{i} w_{ut}^{(i)} \mathcal{R}_{lut}^{(i)}$$

・場所別傾向の俯瞰

多くのユーザに共通する，各場所の習慣性傾向を地図上に可視化する俯瞰表示により推薦内容決定を支援する．ここでは，全ユーザによる同一場所への全訪問の習慣度を平均したものを，赤色ほど習慣度が低い（非習慣），青色ほど習慣度が高い傾向に対応させて，地図上に描画する．

・ユーザ別傾向の縮約

個別ユーザの長期間のログにおける習慣度の変遷を，横軸を日付，縦軸を場所（訪問順），色を習慣度として描画する．加えて，詳細観察対象ユーザの選択を支援するために，複数ユーザの習慣度変遷を羅列した表示も行う．

3.2 可視化例

チェックインログ[1]を用いて位置情報ログの習慣度を用いた傾向観察を行った．

・場所別傾向の俯瞰表示例

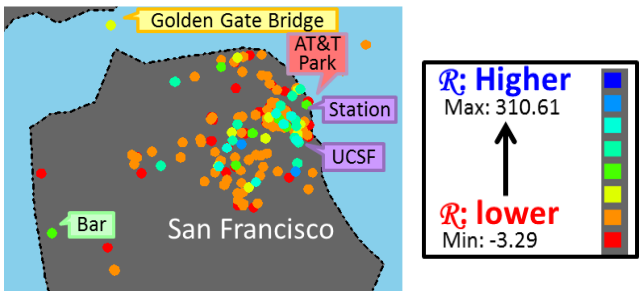

図 4　サンフランシスコにおける場所別傾向俯瞰例

図 4 はサンフランシスコにおける場所別の習慣度平均俯瞰表示結果である．駅(Station)や大学(UCSF)では習慣度が高く，野球場(AT&T Park)や観光名所であるゴールデンゲートブリッジでは習慣度が低い．新しくサンフランシスコに訪れた観光客にはゴールデンゲートブリッジや野球場の情報を，大学や駅をいつもより非習慣的に訪れたユーザには緑色のバーの情報を配信するなどの活用が考えられる．

・特定ユーザの縮約表示例

図 5(左)は，ある 1 ユーザの習慣度変遷の縮約表示例である．一定期間習慣状態や非習慣状態が継続していることが読み取れる．習慣度の継続性などの詳細な議論については，博士論文または文献[5]を参照されたい．また，青色で示される習慣的な訪問が多く見られる場所においても，緑色で示されるやや非習慣的な訪問が見られることが分かる．よく行く場所においても，残業や寄り道などによる軽い非習慣状態を検出することにより，新規店舗への集客が見込める可能性があるため，こういった習慣度変遷の観察は有意義であると期待される．

図 5(右)は，観察対象のユーザの選択を支援するために，複数ユーザの習慣度変遷を羅列した例である．俯瞰・縮約の中間的表示となっており，各セルの背景は各ユーザの習慣度平均を示している．あわせて，ユーザ別の習慣度のばらつきも数値で提示されており，非習慣状態が多いユーザに絞って観察を行うことなどが可能になる．

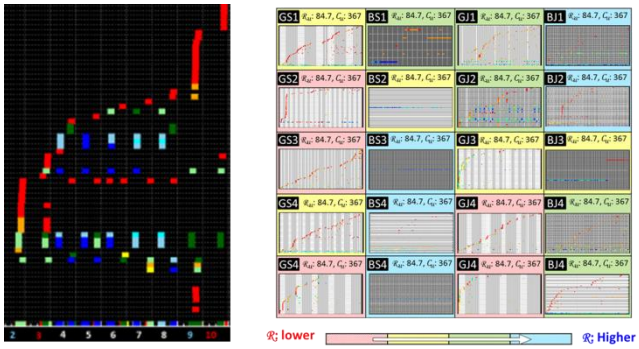

図 5　（左）1 ユーザの習慣度変遷縮約例
（右）複数ユーザの習慣度変遷羅列例

4. クラシック音楽構造の俯瞰・縮約

4.1 提案手法概要

本章では，より特徴的なデータへの対応を目指し，楽譜情報における音楽構造の俯瞰・縮約手法 Colorscore を提案する．楽譜情報の中でも特にオーケストラ楽曲のスコアは段数が多く，楽曲理解は難しい．Colorscore は，各フレーズが担う主旋律，伴奏などの役割を半自動的に抽出することにより，作曲者や編曲者，演奏者による，全体像把握と他編成へのアレンジを支援する．本手法では楽譜データに MIDI を用いる．MIDI では，楽譜上の各パートの各音符について，発音のタイミング，音程，強さなどが数値で記述されている．

・音楽構造の分析

まず，役割判定のためのパターンを付与する．ここでいう役割とは，主旋律，伴奏などである．パターンの例を図 6 に示す．本手法におけるパターンとは，1 パートで構成される十数小節以下の短い楽譜のことであり，本手法では経験者によって MIDI 形式で付与されるものとする．

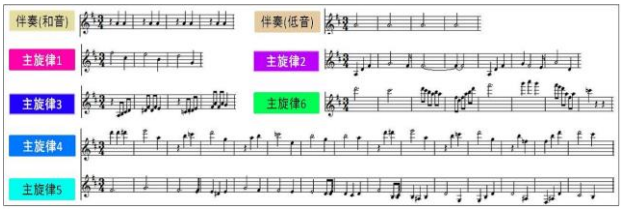

図 6　付与する役割パターンの例

続いて大まかな初期ブロックを生成したのち，与えたパターンと各ブロックをマッチングさせ，各ブロックの役割を反復的に判定する．本手法におけるブロックという概念は，一般的にフレーズに該当し，同パートの連続した音符で構成さ

[1] http://snap.stanford.edu/data/#locnet

れる．ブロックとパターンの類似度算出には，文献[6]のRhythmicActivity(発音タイミングの特徴量)とMelodicActivity(音程遷移の特徴量)をもとに，各ブロックと各パターンのメロディとの距離をコサイン類似度によって求める．特徴量算出方法や分析アルゴリズムの詳細については，博士論文または文献[7]を参照されたい．

・縦横両方向の縮約表示

縦方向の縮約表示では，ユーザの指定した段数にスコアを縮約する．この際，元の楽曲で隣接している数段の中から，主旋律，伴奏，その他の順で，よりパターンとの類似度が高いブロックを選び，それを縮約結果の各段に割り当てる．縮約結果は，MIDI 形式で出力可能なため，効率的な編曲支援ができると期待される．

横方向の縮約表示では，より簡潔に音楽構造を可視化するために，各パートやその役割が前の小節から変化しているかに着目し，縮約表示する．変化のない小節を縮約して表示することで，全体の長さを短く見せることができる

4.2 可視化例

・俯瞰表示

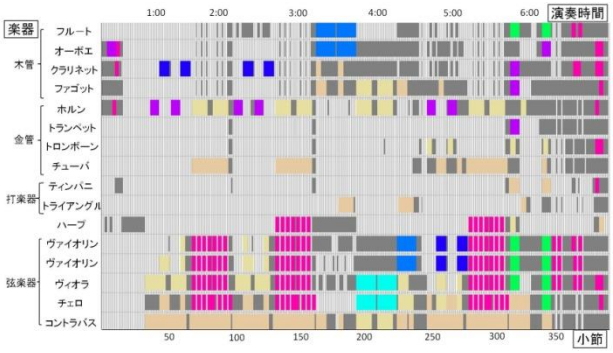

図 7　提案手法の可視化(俯瞰表示)結果例

チャイコフスキーの「花のワルツ」を用いた曲全体の俯瞰表示結果を図 7 に示す．役割パターンは図 6 のとおりである．元の楽譜の段数は 16 段で，ページ数は 33 である．この可視化結果から，クラシック楽曲において典型的な，2 つの主旋律が変奏を伴って繰り返され，音楽構造を構成する様子を把握することができる．また，後半に向けて登場するパート数が増えていく様子や，類似した主旋律が演奏されていてもそれを担うパートが変化する様子，徐々に対旋律が加わる様子も観察することができる．

・縦方向の縮約表示

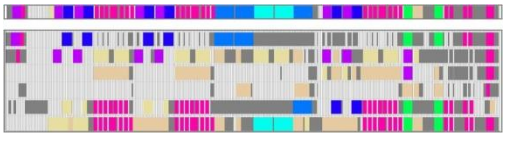

図 8　(上) 1 段の縮約結果 (下) 6 段の縮約結果

図 8 は同楽曲全小節を 1 段，6 段に縮約した結果である．音楽構造を保持して縮約が行われており，編曲等に役立つと期待される．横方向の縮約については紙面の都合で割愛する．

5. まとめ

時系列データの俯瞰と縮約のための 3 つの可視化手法を提案した．3 手法の構築を通し，可視化結果の可読性向上と，縮約箇所選択支援の両方の場面で，提案手法が導入した自動抽出指標が有用であることが分かった．時系列データの可視化分析システムを構築する際には，観察目的とデータ特徴・時間的周期性を考慮した自動抽出指標を導入した上で，時間軸，項目の両方の観点から俯瞰・縮約方法を決定することが有効であると考えられる．

[謝辞]

共同研究者であるお茶の水女子大学 伊藤貴之教授，筑波大学 松原正樹特任助教，明治大学 中村聡史准教授，日本電信電話株式会社 澤田宏主幹研究員および松林達史主任研究員に感謝いたします．また，システムの俯瞰・縮約に関して決済情報をご提供いただいた株式会社インテリジェントウェイブ様に感謝します．

[文献]

[1] M. Krstajić, E. Bertini and D. A. Keim, CloudLines: Compact Display of Event Episodes in Multiple Time-Series, IEEE Transaction on Visualization and Computer Graphics, 17(12) (2011).

[2] J. Zhao, F. Chevalier, E. Pietriga and R. Balakrishnan, Exploratory Analysis of Time-Series with ChronoLenses, IEEE Transactions on Visualization and Computer Graphics, Vol. 17, No. 12 (2011).

[3] L. Berry, T. Munzner, BinX: Dynamic Exploration of Time Series Datasets Across Aggregation Levels, IEEE Symposium on Information Visualization (2004).

[4] A. Hayashi, T. Itoh and S. Nakamura, A Visual Analytics Tool for System Logs Adopting Variable Recommendation and Feature-Based Filtering, The journal of the Society For Art and Science, Volume.13, No.3, pp. 185-197 (2014).

[5] 林亜紀，松林達史，澤田宏，位置情報を利用した情報配信のための習慣度算出手法，日本データベース学会和文論文誌, Vol.13-J, No.1, pp. 64-71 (2014).

[6] 松原正樹，岡本紘幸，佐野智久，鈴木宏哉，延澤志保，斎藤博昭， ScoreIlluminator：スコア色付けによるオーケストラスコアリーディング支援システム，情報処理学会論文誌, Vol. 50, No. 12, pp. 1-12 (2009).

[7] Aki Hayashi, Masaki Matsubara, Takayuki Itoh, Colorscore: Visualization and Condensation of Structure of Classical Music-, Knowledge Visualization Currents: From Text to Art to Culture, Springer-Verlag London, pp. 113-128 (2012).

林 亜紀 Aki HAYASHI

2010 年お茶の水女子大学理学部情報科学科卒業．2012 年同大学院人間文化創成科学研究科理学専攻情報科学コース博士前期課程修了．同年，日本電信電話株式会社入社．2015 年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情報科学領域博士後期課程修了．博士（理学）．時系列データのデータマイニング・可視化の研究開発に従事．日本データベース学会，音響学会各会員．