「データ宝石箱」
～ビジュアルデータマイニングの実現に向けて～

伊藤 貴之　山口 裕美

日本アイ・ビー・エム(株) 東京基礎研究所　E-mail: {itot,yyumi}@trl.ibm.com

コンピュータ・グラフィックス（CG）の研究成果において、「ビジュアルデータマイニング」というキーワードが最近多用されている。本報告ではこのキーワードの定義を、「データに潜在する興味深い現象を、視覚的に発見しやすくするための表現技術」と定義する。

一方筆者らは、階層型データ視覚化手法「データ宝石箱」を提案している。「データ宝石箱」は、階層型データ全体を一画面に展開して表示することで、データの分布を一目で理解するような視覚化手法はないものか、ちょうど宝石店のショーケースのようにデータ全体を見渡せる視覚化手法はないものか？ という発想から生まれた視覚化手法である。筆者らは、ウェブのアクセス傾向の視覚化をはじめとして、多くの題材に「データ宝石箱」を適用している。

本報告では、ビジュアルデータマイニングを実現する諸手法を紹介し、続いて「データ宝石箱」の技術的概要と、ウェブアクセスログに適用した視覚化事例を紹介する。さらに、「データ宝石箱」によって、どのようにビジュアルデータマイニングを実現できるか、について考察し、今後の展望について述べる。

Data Jewelry-Box:
for the Realization of Visual Data Mining

Takayuki ITOH　Yumi YAMAGUCHI

IBM Research, Tokyo Research Laboratory

"Visual Data Mining" is a hot keyword in recent studies in computer graphics area. This report defines the keyword as computer graphics technologies that help to visually find interesting and subconscious trends in given data.

Authors have proposed a hierarchical data visualization technique, "Data Jewelry Box". It represents whole the given data in one display space, so that it provides views to look over whole the data, as if showcase of jewelry shops shows whole the jewelries. Authors have applied the technique to various data including the distribution of accesses of Web sites.

This report introduces some visual data mining techniques. It then introduces the technical overview of Data Jewelry Box, and its application to the visualization of access logs of Web sites. It then discusses how Data Jewelry Box realizes visual data mining, and finally our future works.

1. はじめに

ビジュアリゼーション（可視化・視覚化）は、データの持つ特徴を直感的に理解することを目的とした画面表示技術であり、データ処理技術、コンピュータ・グラフィックス(CG)技術、ユーザーインタフェース技術、などを組み合わせた複合的な研究分野である。これらの技術のうち CG 技術は、データの構成要素に形状や位置などの幾何特性を与え、色や明るさなどの光学特性を与える処理を担当する。この幾何特性や光学特性は、データの特徴を直感的に理解するためのキーポイントとなる。近年では、視覚化のための CG 技術に関する研究成果として、何らかの知見を幾何特性や光学特性に与えることで、データに潜在する興味深い現象を視覚的に発見しやすくする技術が多く報告されている。本報告では、この技術を「ビジュアルデータマイニング」と定義する。

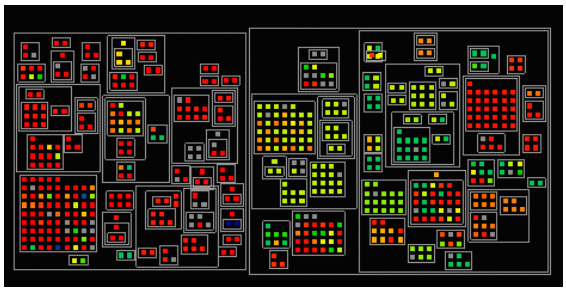

図 1 「データ宝石箱」による階層型データの表示例。

一方筆者らは、階層型データ視覚化手法「データ宝石箱」を提案している[Ito01][Yam03a]。「データ宝石箱」は、階層型データ全体を一画面に展開して表示することで、データの分布を一目で理解するような視覚化手法はないものか、ちょうど宝石店のショーケースのようにデータ全体を見渡せる視覚化手法はないものか？ という発想から生まれた視覚化手法である。

「データ宝石箱」では、葉ノードをアイコンで表現し、その上位階層に相当する枝ノードを長方形の枠で表現する（図 1 参照）。処理手順としては、まず葉ノードを画面配置し、その上位階層の枝ノードを表現する長方形の枠で葉ノードを囲む。続いて、さらに上位階層に着目して、長方形の集合を囲む長方形を作成する。この処理を下位階層から最上位階層に向かって反復することで、データを画面空間に配置する。画面空間を有効活用するために、本手法では長方形をできるだけ隙間なく配置して占有空間の最小化を図る。筆者らは、ウェブのアクセス傾向の視覚化 [Yam03a]をはじめとして、多くの題材に「データ宝石箱」を適用している。

本報告では、まず 2 章にて、ビジュアルデータマイニングを実現する諸手法を紹介する。続いて第 3 章にて「データ宝石箱」の技術的概要を紹介し、また「データ宝石箱」による視覚化結果を紹介する。この結果を参照して、「データ宝石箱」によって、どのようにビジュアルデータマイニングを実現できるか、について考察する。

2. ビジュアルデータマイニング

ビジュアルデータマイニングという単語をウェブの検索エンジンで探すと、概ね以下の 2 種類の意味で使っている事例が多いようである。

(a) データマイニング結果をビジュアルに提示すること。

(b)ビジュアルな技術によって、データマイニング結果に類似した知見を得ること。

本報告では、ビジュアルデータマイニングというキーワードを、(b)の意味で定義する。本章では、視覚化のための CG 技術に着目し、入力データに幾何特性や光学特性を与える過程で、何らかの知見を与えることで、データの視覚的理解を助ける技術を紹介する。

2.1 サイエンティフィック・ビジュアリゼーションにおけるビジュアルデータマイニング

科学技術計算や工業製品解析などの分野においては、計算力学による解析結果を 3 次元離散データで保持することが一般的である。また、医療や気象・地学などの測定結果も、同様に 3 次元離散データとして保持することが多い。このような離散データを対象とした視覚化技術を、サイエンティフィック・ビジュアリゼーションと呼ぶ。この視覚化技術を支える CG 技術は、以下の 2 種類に大別される。

手段 1: ダイレクトアプローチ

離散データ全体を表示対象として、離散データを半透明な光学特性をもつ仮想物質に変換し、これを CG 表示する。代表的な手法として、ボリュームレンダリングがあげられる。

ボリュームレンダリングでは、離散データの数値を光学特性に変換する伝達関数が画像生成結果を支配する。ボリュームレンダリングを用いてビジュアルデータマイニングを実現する有力な手段として、データ中の重要な意味をもつ部位を強調表示するような伝達関数を自動設定することがあげられる。

藤代らは、離散データの数値分布の位相を解析し、その位相からデータ中の重要な数値を予測し、その数値をもつ部位を強調表示するような手法を提案している[Fuj00]。著者らは、藤代らの手法を拡張して、複数の伝達関数を重ね合わせる半自動的な手法を提案している[Yam02]。

これらの手法は、データ中の重要な特徴を理解しやすいような画像表現を実現しているという点で、ボリュームレンダリングのためのビジュアルデータマイニングであると考えられる。

手段 2：インダイレクトアプローチ

離散データの中から、ある特定の数値をもつ部位や、ある条件を満たす部位だけを抽出し、それを曲線や曲面などの幾何形状に変換してグラフィックス表示する。代表的な手法として、等値面や流線があげられる。これらの手法においてビジュアルデータマイニングを実現するためには、データ中の重要な意味をもつ部位を理解できるような幾何形状を自動生成できる必要がある。

小山田らは、科学技術計算結果として得られるベクタ場に対して、渦中心点などの特異点の周辺には興味深い現象が見られることが多いという知見に基づき、特異点を出発点として流線を生成する手法を提案している[Koy98]。この手法は、データ中のベクタ場の特徴的な部位を強調した画像表現を実現するという点で、ベクタ場の視覚化のためのビジュアルデータマイニングであると考えられる。

2.2 インフォメーション・ビジュアリゼーションにおけるビジュアルデータマイニング

3 次元離散データのような、実世界の座標系を持つデータに限定せず、一般的な情報を対象とした視覚化技術を、インフォメーション・ビジュアリゼーション（情報視覚化）と呼ぶ。

サイエンティフィック・ビジュアリゼーションでは、データの構造、性質、用途がある程度確定しており、ビジュアルデータマイニングの研究も系統的に進んでいるのに対して、インフォメーション・ビジュアリゼーションでは、データの構造、性質、用途が非常に多岐にわたっており、その研究は拡散する傾向にあると考える。代表的なサーベイ論文 [Kei02] を見ても、データ構造には 1 次元、2 次元、多次元、テキスト、木構造、グラフ構造、アルゴリズムなどの多岐にわたり、その CG 表示技術やユーザー操作技術も多岐にわたっていることを紹介している。ビジュアルデータマイニングを提唱している典型的な論文 [Rog96] でも、一つのデータに対して多種類のビューを与え、その中から興味深い現象の見られるビューをユーザーに選択させる、というような方法がとられている。

3. データ宝石箱

筆者らは、階層型データを対象とした情報視覚化手法「データ宝石箱」を提案している[Ito01][Yam03a][ZDnet]。本章では、「データ宝石箱」の技術的概要と、ウェブアクセスログの視覚化への応用事例を紹介する。さらに、「データ宝石箱」がどのようにビジュアルデータマイニングに貢献できるか、について考察する。

3.1 技術的概要

1 章で述べた通り、筆者らが提案している階層型データ視覚化手法「データ宝石箱」は、葉ノードをアイコンで表現し、枝ノードを入れ子状の長方形の枠で表現している。

ここで「データ宝石箱」が対象としている一般的な階層型データは、座標情報を持たない。よって図 1 のようなデータ視覚化を実現するためには、データを構成するノードに画面空間上の座標値を与え、データを画面空間に配置するアルゴリズムが必要である。

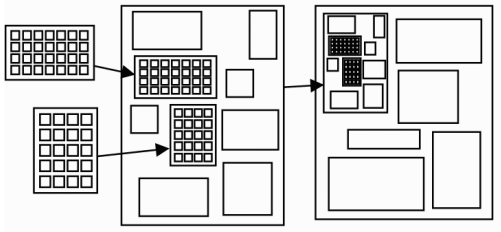

図 2　階層型データの画面配置順。まず最下位階層の葉ノードを配置し、続いて下位階層から上位階層に向かって配置処理を反復する。

図 2 に、「データ宝石箱」による階層型データの画面配置アルゴリズムを示す。本手法では、まず最下位階層に属する葉ノードに対応するアイコン（図 2 の場合は正方形）を隙間無く配置する。続いて、この上位階層に属する枝ノードを表現するために、アイコンを包括する長方形を生成する。さらに、上位階層の枝ノードを表現する長方形群を隙間無く配置し、同様にこれを包括する長方形を生成する。以上の処理を、最下位階層から最上位階層に向けて反復することで、データ全体の配置を決定する。

本手法では、1 個の枝ノードを表現する長方形の枠の内部に、複数の葉ノード(アイコン)や枝ノード(長方形の枠)が配置される。これらのノードを長方形の集合であるとすると、本手法を実現するためには、1 階層を構成する長方形の集合を、以下の条件を満たすように画面空間に配置する技術が必要である。

[条件 1] 長方形どうしが重なってしまうと、データの視覚的理解を妨げるので、隣接長方形と重ならないように長方形を配置する。

[条件 2a] 配置結果の占有面積が大きくなると、それだけ大きなディスプレイ領域を要するので、占有面積を拡大しないように長方形を配置する。

[条件 2b] やむを得ず配置結果の占有面積を拡大するときは、できるだけ占有面積の拡大量が小さい位置に長方形を配置する。また、できる

だけ好ましい縦横比の占有領域を構成するように長方形を配置する。

このような条件を満たすように形状データを配置する問題は、「占有面積の最小化問題」として、VLSI 回路の基板配置、板金や服飾型紙への部品配置、などの用途で知られている。これらの用途では、遺伝子アルゴリズムなどの最適化手法を用いて部品の配置を実現している例が多い。しかし最適化手法には、数分～数時間の計算時間を要する事例が多く、対話的操作を要する視覚化の分野には向かない。「データ宝石箱」では、占有面積が最小でなくてもいいから、ある程度良好な結果を短時間に算出する配置手法を用いる。文献[Ito01][Yam03a]では、その配置手法の一例として、長方形群を隙間なく配置する高速な新しいアルゴリズムを提案している。

なお、[条件 2b]における「好ましい縦横比」とは、最上位階層においてはディスプレイやウィンドウの縦横比、それ以外の階層においては縦:横=1:1 であるとする。

3.2 ウェブアクセスログの視覚化事例

筆者らは文献[Yam03a][ZDnet]にて、「データ宝石箱」をウェブアクセスログの視覚化に適用した事例を紹介している。この事例では、ウェブサーバーに蓄積されるアクセスログファイルを入力データとする。このとき筆者らの実装では、以下の 2 つのビューを自動表示する。

(1) サイトマップ: アクセスログに記述された URL から、ウェブページ群をディレクトリ階層に基づいて階層型データに整理し、「データ宝石箱」を用いてそれを画面配置したもの(図 3(b)参照)。
(2) 統計グラフ: ユーザーがアクセスログ中の 1 個の属性を指定した時に、その属性に基づいてアクセス数を集計し、その結果を棒グラフで表示したもの(図 3(c)参照)。

さらに筆者らの実装では、サイトマップと統計グラフとの間に以下の 2 種類の連携操作機能を提供する。

[連携 1] 統計グラフからサイトマップへの反映

ユーザーが統計グラフの 1 箇所をクリックすると、本手法はクリックされた箇所に該当するアクセス数をウェブページごとに集計する。そして、個々のウェブページのアクセス数に応じて、アイコンに高さを与える。このようにして、特定の属性値をもつアクセスの分布を、サイトマップ上で視覚化することができる（図 3(d)参照)。

[連携 2] サイトマップから統計グラフへの反映

ユーザーがサイトマップ上で関心のあるウェブページのアイコンをクリックすると、そのウェブページへのアクセスを集計した統計グラフが表示される。これによりユーザーは、サイト全体のアクセス傾向だけでなく、関心のある特定のウェブページに対するアクセス傾向も知ることができる（図 3(e)参照)。

ここで、筆者らの実装を用いて、実在するウェブサイトの 1 週間のアクセスログを視覚化した実験例を示す。筆者らはまず、アクセスを日付で分類集計して 7 本の棒グラフを作成し、さらに各項目を 1 時間単位で分割することで棒グラフを 24 色に色分けした。以上の分類にしたがって、横軸が日付、縦軸がアクセス数を示す統計グラフを作成した（図 3(c)参照)。これを見ると、最終日のアクセス数が他の日に比べて極端に高いことがわかった。ここでまず[連携 1]を用いて、統計グラフ上で、最終日のある 1 時間をクリックし、サイトマップ上でその 1 時間のアクセス分布を表示した(図 3(d)参照)。

このとき、図 3(d)中の右下部にある、四角で囲んだアイコンが表すページが、午前中に突出して多くのアクセス数をもつことがわかった。[連携 2]を用いて、そのページへのアクセスを対象としてリンク元の URL で分類した統計グラフを表示した。すると、ある新聞会社のオンラインニュースの URL から訪れているサイト閲覧者が多いことがわかった(図 3(e))。リンク元であるオンラインニュースにアクセスしてみると、そこにアクセス数の多かったページが取り上げられていたことがわかった。以上の結果から、午前中に新聞サイトを見て、そのリンクをたどってこのページに来た人が多かったことを推測した。

また、図 3(d)中の右上部にある、丸く囲んだ長方形が表すディレクトリ中のほとんどのページが、1時間以内にアクセスされていたことがわかった。続いて[連携 2]を用いて、これらのページへのアクセスをサイト閲覧者の IP アドレスで分類して表示すると、すべてのページに同一 IP アドレスからのアクセスがあることがわかった。これらの結果から、あるディレクトリのファイルをすべて見ている熱心なサイト閲覧者が存在していたことがわかった。

3.3 「データ宝石箱」とビジュアルデータマイニング

3.2 節で紹介したウェブアクセスログの視覚化事例について、「データ宝石箱」がどのように貢献したのか、について考察する。

既存の市販ウェブアクセス分析ツールの多くは、まず棒グラフや折れ線グラフ、ランキング表などの単純な表示を用いて、アクセスの非常に概略的な統計結果を提示し、続いてユーザー操作によって選択的にアクセス傾向を探索する、というように構成されている。それに対して、図 4 に示した表示例では、2000 以上のウェ

ブページをもつサイトのアクセス分布を一画面に全て表示することで、その中のごく局所的な1ページ、あるいは1ディレクトリに関する興味深い傾向を「最初の一目で」発見させることに成功している。
つまり「データ宝石箱」では、データ全体を一画面にすべて表示させるというコンセプトにより、データ全体にわたる概要を提示しているだけでなく、その中のごく局所的な部分に見られる潜在的な現象を、最初の一目で発見させる役割をも同時に果たしていると言える。いずれにしても、データの理解を助けるための CG 技術という意味で、「データ宝石箱」はビジュアルデータマイニングのための一手法であると言うことができるだろう。

4. 今後の展望

本報告では、ビジュアルデータマイニングというキーワードを定義し、CG 技術の立場からその実現を目指す諸手法を紹介した。また、筆者らの情報視覚化手法である「データ宝石箱」を紹介し、それがどのようにビジュアルデータマイニングに貢献できるかという点について考察した。
今後の筆者らの展望について、以下の 2 点から述べたい。

4.1 他のデータを用いた実証

筆者らはすでに、多くの研究グループとの協力により、以下の題材について議論を始めており、またいくつかについては実験段階まで到達している。

(1) 非常に汎用性の高いデータ。例えば以下のようなデータを対象としている。
- ファイルシステムの階層構造の表示。
- 大量の文書データ群から得られるキーワード群の分布図表示。

(2) アクセスログ以外の題材を用いたウェブサイトの視覚化。例えば以下のような観点からのウェブサイトの視覚化を考えている。
- 検索エンジンの抽出結果として得られるウェブページ群の分布図表示。
- 個々のウェブページのデザインがサイトのポリシーを満たしているか、などの検閲結果に対する視覚化。

(3) リアルタイム性の高いデータに対する監視目的での視覚化。例えば以下のようなデータについて実験環境を構築したい。
- 分散計算環境で稼動するプロセス群の分布図表示[Yam03b]。
- ネットワークへのハッキング行為の発見のための監視表示。

(4) 専門性の高いデータを一般人に直感的に理解させるための視覚化。例えば以下のような用途を考察中である。
- 科学技術計算結果のデータベース内容の直感的なプレゼンテーション。

以上の多様な視覚化を通して、「データ宝石箱」がどのようにビジュアルデータマイニングを実現できたか考察し、本報告の続報としたい。

4.2 CG 技術としての拡張

「データ宝石箱」は、大量のデータを、省略せずに全て一画面に格納表示することで、ビジュアルデータマイニングにむけて一定の成果をあげることができた。しかし一方で、「大量のデータを全て表示する」というポリシーが逆効果を生むこともありえる。例えば、視覚的な情報量が多すぎて、却って理解を妨げる、という問題が生じる可能性がある。この問題に対して既存の情報視覚化の諸手法は、
- CG 表示前のデータ処理の過程で、情報量を適正化する。
- CG 表示後のユーザー操作によって、情報を選択しながら表示する。

というような解決方法をとっていることが多い。以下、CG 技術の観点からどのようにこの問題を解決できるか、について展望を述べる。
CG 技術は視覚化の他に、科学技術や工業技術の支援、アートやエンターテインメント、などの目的で発展してきた。近年ではアート系の CG 技術の一環として、情報芸術（インフォマティック・アート）という研究分野が発展している。これは CG 技術による美しい表現が、一般人にも親しみやすい情報提示を可能にする、という考え方に基づく研究分野である。
「データ宝石箱」は、大量のデータを素直に全部そのまま表示することを目的としている。いわば、数千本、数万本の樹木を素直に全部描いた 19 世紀以前の絵画技法のような技術である。それに対して近年の CG 技術では、20 世紀以降の印象派芸術のように、風景の概略的な傾向だけを表現した画像生成技術や、ポップアートのように、親しみやすさを前面に出して非忠実にシーンを表現する画像生成技術が研究されている。このような考え方を適用することで、現状とは異なる情報提示が可能になるのではないか、という点を考察したい。
また「データ宝石箱」は、直交 2 次元座標系にデータ群を配置し、それに直交する 3 個目の次元で統計値を表す、という意味では、幾何処理の観点からも素直な手法である。それに対して近年の CG 技術では、あえて歪んだ空間を仮想してデータを配置することにより、視覚効果の高い多視点画像や、エッシャーのだまし絵の

ような非現実的な画像を生成する研究が進んでいる。これらの考え方を適用することで、現状よりも印象の高い情報提示が可能になるのではないか、という点を考察したい。

謝辞

先行研究に関して貴重なご教示をいただいた、お茶の水女子大学藤代一成教授に感謝します。また、日頃討論いただく日本アイ・ビー・エム(株)東京基礎研究所松澤裕史研究員、長野徹研究員、他多くの研究員に感謝します。また、「データ宝石箱」の今後の展望に関してご意見をいただいた、岩手県立大学土井章男教授、京都大学小山田耕二助教授、北陸先端科学技術大学院大学宮田一乘教授に感謝します。

参考文献

[Fuj00] Fujishiro I., Azuma T., Takeshima Y., Takahashi S., Volume Data Mining Using 3D Field Topology Analysis, IEEE Computer Graphics & Applications, Vol. 20, No. 5, pp. 46-51, 2000.

[Ito01] 伊藤,梶永,池端,データ宝石箱:大規模階層型データのグラフィックスショーケース, 情報処理学会グラフィクス&CAD 研究会, 2001-CG-104, 2001.

[Kei02] Keim D. A., Information Visualization and Visual Data Mining, IEEE Trans. On Visualization and Computer Graphics, Vol. 8. No. 1, pp. 1-8, 2002.

[Koy98] Koyamada K. and Itoh T., Seed Specification for Displaying a streamline in an Irregular Volume, Engineering with Computer, Vol. 14, pp. 73-80, 1998.

[Rog96] Rogowitz B. E., Rabenhorst D. A., Gerth J. A., Kalin E. B., Visual Cues for Data Mining, SPIE/SPSE Symposium, pp. 275-301, 1996.

[Yam02] 山口, 藤代, 竹島, 高橋, 伊藤, ボリュームデータマイニングのための伝達関数の合成, 映像情報メディア学会論文誌, Vol. 56, No. 6, pp. 973-978, 2002.

[Yam03a] 山口, 伊藤, 池端, 梶永, 階層型データ視覚化手法「データ宝石箱」とウェブサイトの視覚化, 画像電子学会論文誌ビジュアルコンピューティング特集号, 査読中.

[Yam03b] Yamaguchi Y., Itoh T., Visualization of Distributed Processes Using “Data Jewelry Box” Algorithm, CG International 2003, accepted.

[ZDnet]http://www.zdnet.co.jp/news/0210/25/nj00_vc_ibm.html

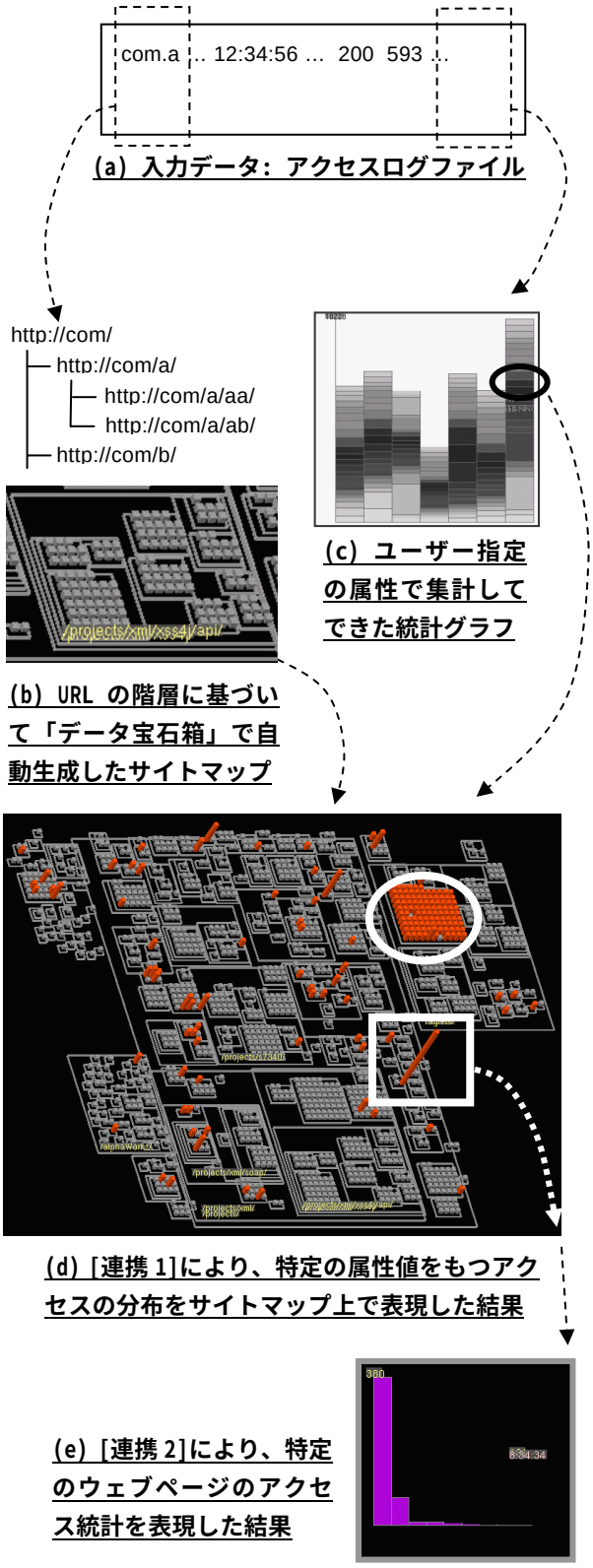

図 3 「データ宝石箱」を用いたウェブアクセスログの視覚化