SRA GROUP
GSLetterNeo Vol.93
2016年4月

カタマリを見つけて辿れるヒストリデータブラウザ(3)

松原 伸人、土屋 正人
matubara@sra.co.jp, m-tsuchi@sra.co.jp

大量のデータをインタラクティブに操作する、Webブラウザ上で動作するアプリケーションを開発しています。

「**カタマリを見つけて辿れるヒストリデータブラウザ**」[1]は、歴史データや開発ログといった大量のイベントを、時間軸上に並べて表示することで、イベントが連続的に起きている時間帯や切れ目を見つけて辿れるようにするWebアプリケーションのプロトタイプです。

このプロトタイピングでは、どのような実装をすれば大量のデータをインタラクティブに扱えるか？ 反対に、実現したいインタラクションを実装してみた結果として、インタラクティブ性能がどうなるか？ ということを試行錯誤して探っています。

◆ “まとまり”をつくる

大量のデータの中から特定のイベントを見つけたり、似たようなイベントを見つけたりするには、1つ1つイベントを見ていって、タグをつけて仕分けていきます。そうすることで、特定のタグが付いているイベントを検索して順に見ていったり、同じタグごとにイベントをまとめて見ることができます。

人手で仕分けた“まとまり”は、どんな“まとまり”なのか分かりやすくなります。しかしながら、人手であるため、時々間違っていたり、仕分けられずに抜けていたりすることもあります。

すべてのイベントの内容が、テキストや画像で表されているなど、データ形式が決まっている場合は、イベントデータからテキストや画像の特徴抽出し、特徴量の比較によってクラスタリングを機械的に行えます。

この場合、人の手作業によるミスや検出漏れの無い仕分けを行えますが、それぞれのクラスタがどんな“まとまり”なのかを理解するのが難しくなります。

◆ クラスタリング結果

図1～9は、京都の歴史データ約12,000の出来事をクラスタリングした結果の各クラスタを色分けし、並べて表示しています。

出来事の類似度の閾値を 9 段階(0.08 0.1 0.12 0.15 0.2 0.3 0.4 0.5 0.7)に変えて行った9種類のクラスタリング結果があります。

図において、縦方向は、上から紀元前のイベントに始まり、下に向かって現代のイベントが並ぶ時間軸です。横方向は、異なるクラスタを色と位置を変えて並べています。クラスタの位置や色は、クラスタのIDを基に決めているため、色の近さや位置に意味はありません。

すべての出来事が必ずどれかのクラスタに入るのではなく、どのクラスタにも入らない出来事もあります。

また、1つの出来事が1つのクラスタに入るだけでなく、複数のクラスタに入ることもあります。

閾値が低いと、どれかのクラスタに入る確率が上がり、0.08 の時はクラスタ数が 1562 と多く、閾値が高い 0.7 だとクラスタに入りにくくなり、結果的にクラスタ数 144 と少なくなるようです。

各出来事の内容を表すテキスト文章を形態素解析した結果を用いて、クラスタリングが行われているようです。

[1] GSLetterNeo Vol.91、92を合わせてご一読頂けると幸いです。

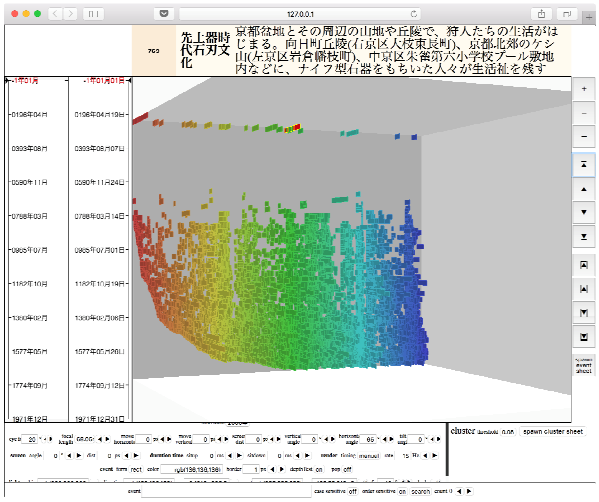

図 1　閾値 0.08
クラスタ数 1562 イベント数 25263

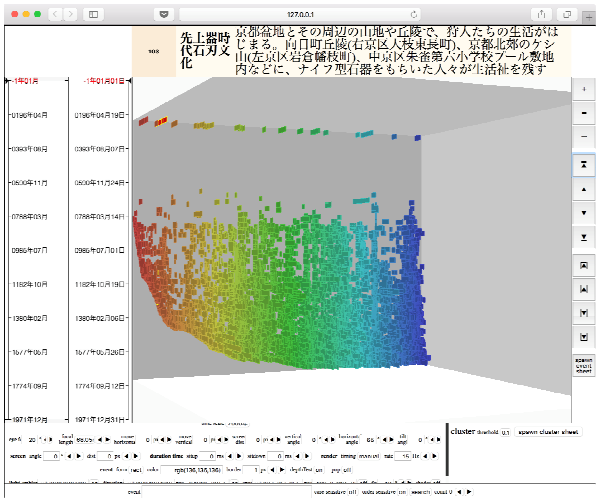

図 2　閾値 0.1
クラスタ数 1463 イベント数 21849

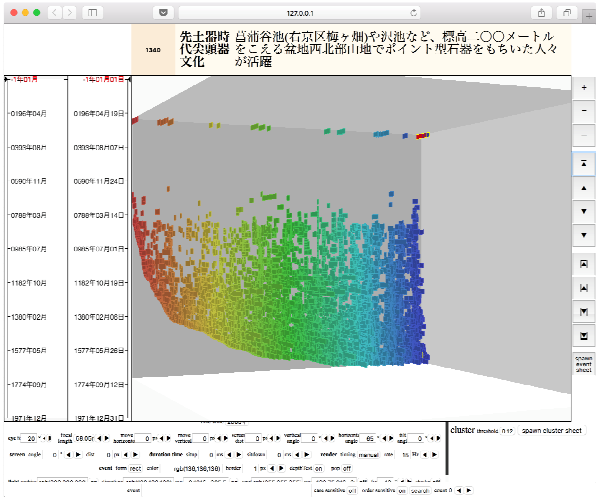

図 3　閾値 0.12
クラスタ数 1366 イベント数 16841

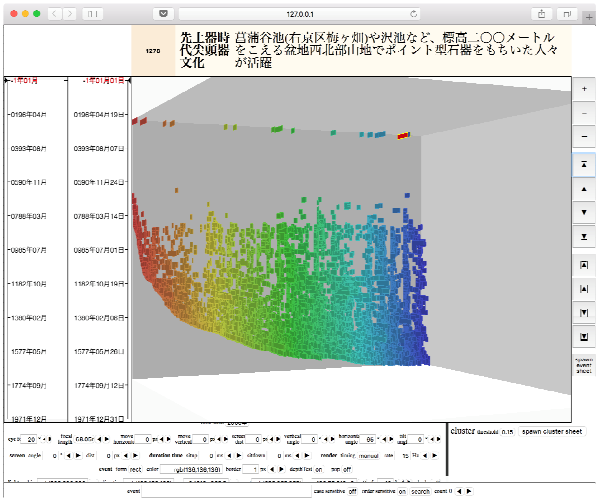

図 4　閾値 0.15
クラスタ数 1409 イベント数 13530

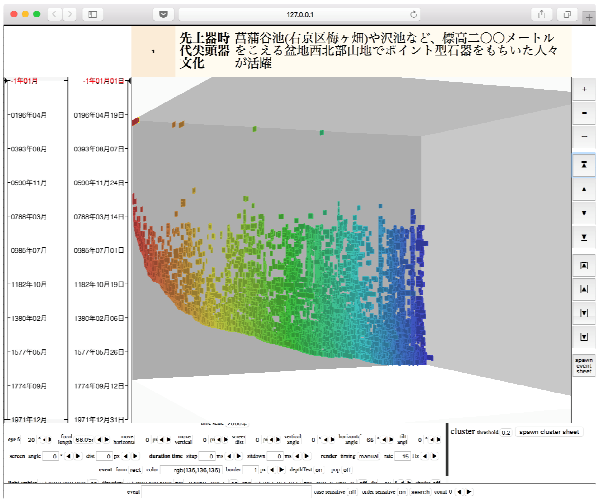

図 5　閾値 0.2
クラスタ数 1368 イベント数 11102

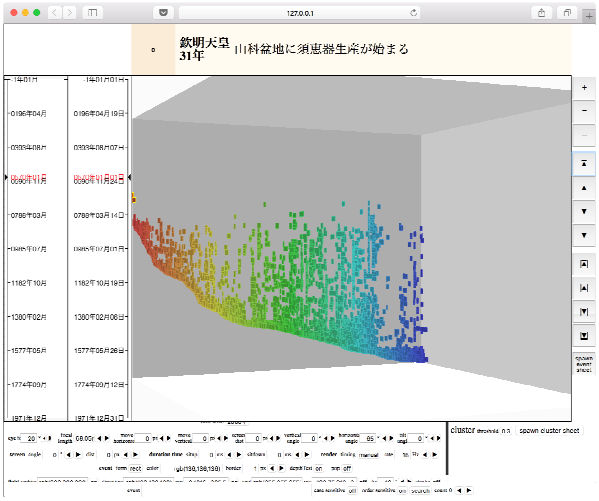

図 6　閾値 0.3
クラスタ数 1061 イベント数 5105

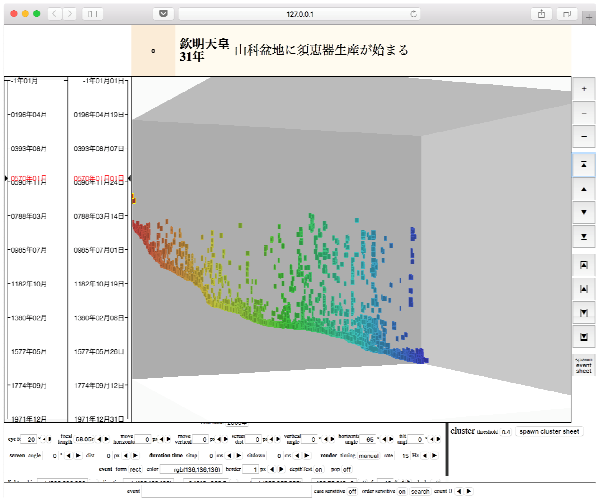

図 7　閾値 0.4
クラスタ数 650 イベント数 2402

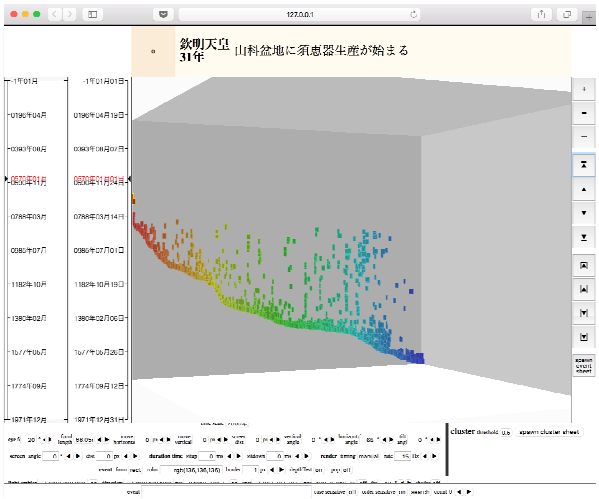

図 8　閾値 0.5
クラスタ数 479 イベント数 1497

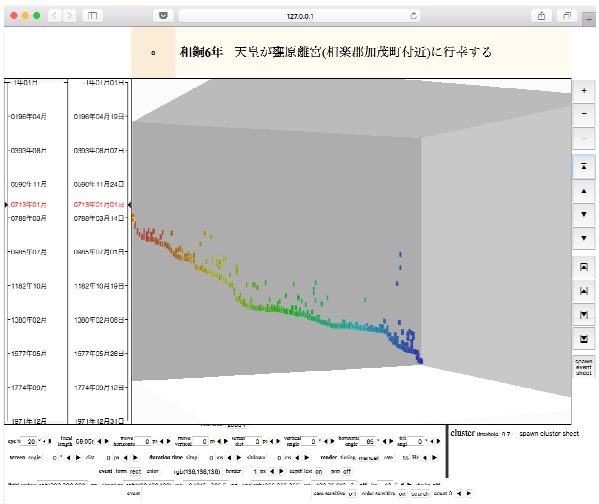

図 9　閾値 0.7
クラスタ数 144 イベント数 343

画面中央のデータエリアでは、選択しているイベントを赤色、選択しているクラスタを黄色で表しています。

選択中のイベントが入っているクラスタ ID のリストとイベントの内容を画面上部に表示します。選択中のクラスタ ID はリスト内に太字で表しています。クラスタリスト内の ID を選ぶと、データエリア上で、選んだクラスタに含まれるイベントを選んで表示します。

画面右端下部のボタンを押して、クラスタ内のイベントを時間順に選択できます。

これらの操作を用い、イベントを選んで、属しているクラスタを見て、各クラスタでのイベントを順に辿り、各クラスタがどういうイベントの集まりなのかを見ていきます。

このようにして、各クラスタがどういうイベントの集まりなのかを見ていきます。

画面最下部のイベント検索ツールは、検索文字列を含むイベントを検出し緑色で表します。検索文字列を空白文字で区切ると、指定した文字列の順の内容を持つイベントを検出します。 ◂ボタンと ▸ボタンを押すと、検索結果を順に選択して表示します(図 10)。

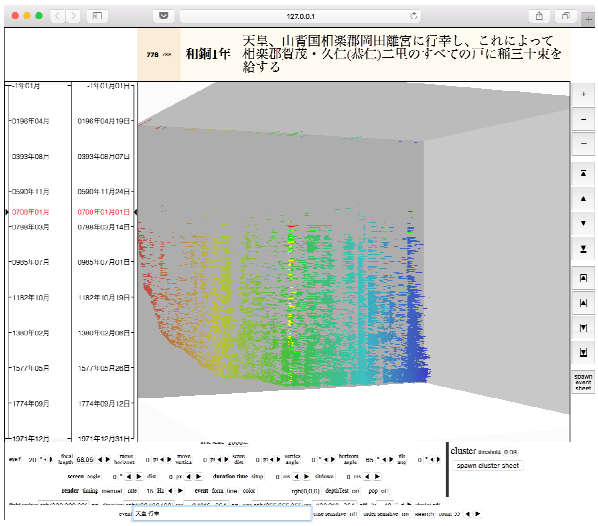

図 10　"天皇 行幸"で検索した画面

イベント内容をテキスト検索できると、クラスタ内のイベントを表す文字列を手掛かりにして、内容が似ているイベントを見つけ、類似するクラスタを探すことができます。

◆ 使用データについて

掲載図で使用しているデータは、「ユーザーの主体的理解醸成のためのデータ表現とインタラクティビティのデザイン」研究グループ（主たる共同研究者：中小路久美代）[2]による研究の一部として、京都大学学際融合教育研究推進センターデザイン学ユニットの北 雄介先生と共同で、京都の歴史年表から収集したテキストデータです。地名や人名を抽出してイベントを分類しています。

2　本研究グループは、JST 戦略的創造推進事業 CREST「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」領域 H26 年度採択課題「データ粒子化による高速高精度な次世代マイニング技術の創出」（代表：宇野毅明（NII 教授））のメンバーです。

GSLetterNeo Vol.93
2016 年 4 月 20 日発行
発行者●株式会社 SRA 先端技術研究所
編集者●土屋正人

バックナンバを公開しています●http://www.sra.co.jp/gsletter
ご感想・お問い合わせはこちらへお願いします●gsneo@sra.co.jp

株式会社SRA
〒171-8513 東京都豊島区南池袋２－３２－８

夢を。Yawaraka Innovation
やわらかいのべーしょん